Robuste kontinuierliche Spracherkennung für mobile Informationssysteme

Zur Erlangung des akademischen Grades
eines Doktors der Ingenieurwissenschaften
der Fakultät für Informatik
an der Universität Karlsruhe (Technische Hochschule)
genehmigte

Dissertation

von

Martin Westphal

aus Belair bei Adelaide (Australien)

Tag der mündlichen Prüfung:	15. Juni 2000
Betreuer:	Prof. Dr. A. Waibel
Korreferent:	Prof. Dr. G. Ruske

Berichte aus der Informatik

Martin Westphal

Robuste kontinuierliche Spracherkennung für mobile Informationssysteme

Shaker Verlag
Aachen 2001

Die Deutsche Bibliothek - CIP-Einheitsaufnahme

Westphal, Martin:
Robuste kontinuierliche Spracherkennung für mobile Informationssysteme / Martin Westphal. Aachen : Shaker, 2001
(Berichte aus der Informatik)
Zugl.: Karlsruhe, Univ., Diss., 2000
ISBN 3-8265-8901-7

Printed in Germany.

ISBN 3-8265-8901-7
ISSN 0945-0807

Shaker Verlag GmbH • Postfach 1290 • 52013 Aachen
Telefon: 02407 / 95 96 - 0 • Telefax: 02407 / 95 96 - 9
Internet: www.shaker.de • eMail: info@shaker.de

Danksagung

Mein erster Dank gilt Prof. Dr. Alex Waibel für die Betreuung dieser Arbeit sowie die zahlreichen wissenschaftlichen Impulse. Durch seine Forderung nach den Sternen zu greifen und die Erwartung das Beste nochmals zu übertreffen, spornte er mich zu höchsten Leistungen an. Ich bedanke mich für die Möglichkeit an einem technisch hervorragend ausgestatteten Institut in einer Gruppe hochkarätiger Wissenschaftler an einem der weltbesten Spracherkenner mitzuwirken. Besonders wertvoll empfand ich meine Aufenthalte an der Carnegie Mellon University Pittsburgh USA, wo sich das zweite Standbein von Prof. Waibels Forschungsgruppe befindet, sowie die zahlreichen Teilnahmen an Konferenzen und Tagungen.

Ebenfalls zu hohem Dank verpflichtet bin ich Prof. Dr. Günther Ruske von der TU München für die Übernahme des Korreferats, das große Interesse mit dem er meiner Arbeit begegnete und ganz besonders für die wertvollen Anregungen, die mir bei ihrer Fertigstellung sehr halfen.

Diese Dissertation entstand während meiner Mitarbeit am Institut für Logik, Komplexität und Deduktionssysteme der Universität Karlsruhe, wo ich bei mehreren Forschungsprojekten für Merkmalsextraktion und akustische Modellierung zuständig war: Switchboard ((D)ARPA), dem Verbundprojekt Verbmobil (BMBF), View4You, VODIS (EU) und TalkingMap (EML). Mein Dank gilt all meinen Kollegen und Mitarbeitern anderer Forschungsgruppen, die meinen Weg in dieser Zeit begleiteten und mir wertvolle Einblicke in die komplexe Welt der Spracherkennung vermittelten.

Für meine Beherbergung in Pittsburgh bedanke ich mich bei Torsten Zeppenfeld wie auch für die vielen Diskussionen, die ich dort sowohl mit ihm als auch mit Michael Finke, Ivica Rogina, Bernhard Suhm und Puming Zhan führte. Nachhaltigen Eindruck hinterließen bei mir die wissenschaftlichen Gespräche mit Prof. Richard Stern.

Teile dieser Dissertation entstanden im Rahmen des VODIS-Projektes. Hier gebührt besonderer Dank meinem Pittsburgher Kollegen Matthias Denecke, der unermüdlich an den Komponenten zum Sprachverstehen feilte, ohne die der hier entwickelte Spracherkenner nur für sich und nicht in einem Gesamtsystem hätte betrachtet werden können. Ebenso Anteil an diesem Projekt hatten meine Karlsruher Kollegen Uwe Meier und Petra Geutner, aus deren Arbeit ich inhaltsreiche Erkenntnisse ziehen konnte. Stets wertvolle Diskussionspartner unter den Kooperationspartnern waren mir Luis Arevalo und Dirk Van Compernolle.

Immerzu eine Quelle guter Ideen waren mir mit ihrer hohen Fachkompetenz meine Karlsruher Kollegen: Jürgen Fritsch, Hermann Hild, Stefan Jäger, Thomas Kemp, Detlef Koll, Stefan Manke, Klaus Ries, Jürgen Reichert, Ivica Rogina, Thomas Schaaf, Florian Schiel, Tanja Schultz, Tilo Sloboda, Hagen Soltau, Rainer Stiefelhagen und nicht zuletzt Monika Woszczyna. Danke auch für die während dieser Zeit geknüpften privaten Freundschaften.

Bei meinen beiden langjährigen Zimmerkollegen Hermann Hild und Tanja Schultz weiß ich ihr immer „offenes Ohr" sehr zu schätzen. Meinem neu hinzugekommenen Zimmerkollegen Christian Fügen, der das laufende Projekt des tragbaren Touristenführers übernahm und mir so etwas „Luft" in der Endphase verschaffte, fühle ich mich sehr verbunden.

Für den anregenden Gedankenaustausch und das Korrekturlesen meiner Arbeit gilt mein herzlicher Dank: Thomas Kemp, Ivica Rogina, Hagen Soltau, Jürgen Reichert, Hermann Hild und besonders Jutta Klein.

Dank möchte ich ebenfalls unserem System-Administrator Frank Dreilich und seinen Vorgängern Markus Baur, Thomas Walther und Martin Klein aussprechen, ohne deren selbstlosen Einsatz viele meiner für diese Arbeit notwendigen Experimente aus technischen Gründen nicht hätten durchgeführt werden können. Danke auch an unsere gute Fee aus dem Sekretariat Silke Dannenmaier, die trotz ihrer vielen Aufgaben an unserem Institut stets für Ordnung und Harmonie sorgte.

Zuletzt und daher auch an bevorzugter Stelle möchte ich mich bei meiner Familie und insbesondere bei meiner lieben Ehefrau Andrea bedanken, die mir so manches durchgearbeitete Wochenende nachsah. Mit zärtlicher Liebe erwähne ich hier auch meine zweieinhalbjährige Tochter Lara Tabea, die mir nach einer kurzen Unterbrechung zum gemeinsamen Malen, Puzzlespielen oder Kneten stets eine Quelle von guter Laune und neuer Schaffenskraft beim Zusammenschreiben dieser Arbeit war. Außerdem gewann ich durch sie interessante Erkenntnisse über die Spracherkennung und das Lernverhalten eines menschlichen neuronalen Netzes, das mit den Randbedingungen: beschränktes, aber sich stetig erweiterndes Vokabular, Aufbau einer Grammatik, kontextabhängiges Unterscheiden sehr ähnlicher Wörter, unterschiedliche Intonation usw. zurechtkommen muss. Und dabei hat ihr bisher niemand beigebracht was ein Phonem ist.

Karlsruhe, den 26. April 2000

Martin Westphal

Inhalt

Robuste kontinuierliche Spracherkennung für mobile Informationssysteme

Mathematische Schreibweise

Signale, Vektoren und Matrizen

$x(t)$	Zeitsignal zum Zeitpunkt t
$X(\omega)$	Spektrum mit Kreisfrequenz ω
$\mathbf{x}$	Vektor
x_i	einzelner Koeffizient von $\mathbf{x}$
$\mathbf{x}[k]$	Merkmalsvektor (Muster) für Analyserahmen k
$x_i[k]$	einzelner Koeffizient von $\mathbf{x}[k]$
$\mathbf{X}$	Matrix, z.B. Folge von Merkmalsvektoren
$\mathbf{x}^T$, $\mathbf{X}^T$	transponierter Vektor, transponierte Matrix
$\lvert\mathbf{x}\rvert$, $\lvert\mathbf{X}\rvert$	Betrag des Vektors, Determinante der Matrix
$tr(\mathbf{X}) = \sum_i x_{ii}$	Spur (engl.: *trace*) der Matrix
$\Delta\mathbf{x}$	Differenzvektor

Oft verwendete Merkmalsräume und Kennzeichnungen von Merkmalen

x^s	im Spektralbereich
x^l	im logarithmierten Spektralbereich (*Log*-Spektrum)
x^c	im cepstralen Bereich
$\tilde{x}$	gestörter Wert
$\hat{x}$	Schätzwert
x^*	optimaler oder besonderer Wert

Wahrscheinlichkeiten und Statistik

$\mathrm{P}(w)$	Wahrscheinlichkeit
$\mathrm{p}(\mathbf{x})$	Wahrscheinlichkeitsdichte
$\mathrm{P}(w \mid ...)$	bedingte Wahrscheinlichkeit
$\hat{\mathrm{P}}(...)$	Schätzwert der Wahrscheinlichkeit
$\mathrm{E}\{f(\mathbf{x})\} = \int f(\mathbf{x}) \cdot \mathrm{p}(\mathbf{x})\, dx$	Erwartungswert
$\mu_i = \bar{x}_i = \frac{1}{N}\sum_{k=1}^{N} x_i[k]$	Mittelwert
$\sigma^2{}_x$, $\sigma_{x,y}$	Varianz von x, Kovarianz von x und y
$\sigma_i^2 = \sigma_{ii}$, σ_{ij}	Varianz, Kovarianz der Koeffizienten i und j
$\mathbf{C}$	Kovarianzmatrix

1 Einführung

Die eigentliche Gefahr in unserem technischen Zeitalter ist nicht so sehr,
dass Maschinen beginnen, wie Menschen zu denken,
sondern dass Menschen wie Maschinen denken.

Kapitelübersicht: Die automatische Spracherkennung hat in den letzten Jahren einen großen Schritt aus den Forschungslabors hinein in reale Anwendungen getan. In diesem einleitenden Kapitel sollen zunächst verschiedene Einsatzmöglichkeiten und deren Problematiken beleuchtet werden. Nach dem ersten, allgemeineren Abschnitt 1.1 wird im nächsten Abschnitt 1.2 auf eine speziellere Art der Anwendung, nämlich **mobile Informationssysteme**, eingegangen. Abschnitt 1.3 vergleicht zwei Möglichkeiten zur Sprachbedienung solcher Systeme. Der hier schließlich bevorzugte Ansatz basiert auf der Erkennung spontaner, kontinuierlich gesprochener Anfragen. Der folgende Abschnitt 1.4 geht detaillierter auf die Schwierigkeiten bezüglich der hierfür verwendeten **kontinuierlichen Spracherkennung** und die möglichen Ursachen ein.

Eine der wichtigsten Voraussetzungen für den sinnvollen Einsatz und daher auch die Hauptthematik der vorliegenden Arbeit ist die **Robustheit** der Erkennung in wechselnden Umgebungen. Die genaue Zielsetzung und einen Überblick über die Arbeit gibt der letzte Abschnitt 1.5 dieses Einführungskapitels.

1.1 Automatische Spracherkennung für reale Anwendungen

Der Mensch in der heutigen Informationsgesellschaft ist stark auf die Unterstützung durch Geräte jeglicher Art angewiesen. Auch wenn diese, vielfach schon mit Mikrorechner ausgestatteten Maschinen zu fast unverzichtbaren Hilfsmitteln geworden sind, ergeben sich besonders bei komplexen Geräten immer wieder Schwierigkeiten bei der Bedienung. Eine der Ursachen liegt in den unterschiedlichen Fähigkeiten von Mensch und elektronischem Rechenhirn, was auf der anderen Seite aber den großen Nutzen für den Menschen begründet. Beispielsweise können heutzutage handliche Geräte schon sehr viele Informationen speichern. Das Auffinden einer gewünschten Information erfordert jedoch eine genaue Spezifikation, die zu formulieren dem ungeübten Benutzer sehr schwer fällt. Von der zwischenmenschlichen Kommunikation sind wir gewöhnt, dem Gegenüber sehr viel einfachere, unvollständige Anfragen zu stellen. Aus seinem Weltwissen und dem augenblicklichen Kontext ist dieser dann meist in der Lage, die fehlenden Stücke zu ergänzen oder er klärt sie mittels Rückfragen.

Wie soll also die Interaktion zwischen Mensch und Maschine gestaltet werden? Das Schlagwort *„benutzerfreundlich"* versucht dieses Ziel, wenn auch nur sehr unscharf, zu beschreiben. Der Aufwand wird dabei möglichst auf die Maschine verlagert. Die Eingabe

eines Computerprogramms mit einer alphanumerischen Tastatur ist sicherlich eine benutzerfreundlichere Eingabeform als die Verwendung von Lochstreifen mit binär dargestellten Befehlen. Für den Computer stellt es aber einen Mehraufwand dar, da die eingegebenen Zeichen angezeigt und in ein Programm übersetzt werden müssen. Der Lernaufwand für den Benutzer reduziert sich dafür deutlich, denn er kann Begriffe und Symbole verwenden, die ihm aus der Alltagssprache und dem Schriftverkehr bekannt sind. Auch diese mussten irgendwann gelernt werden. Da sie aber für die zwischenmenschliche Kommunikation verwendet werden, kann man von ihrer soliden Beherrschung ausgehen.

Das gesprochene Wort ist mittlerweile auch gut maschinell zu verarbeiten. Neben der Vertrautheit für den Menschen, zeichnet es sich durch weitere Vorteile gegenüber anderen Kommunikationsmitteln aus:

- Im Vergleich zu einer alphanumerischen Tastatur ist die Eingabe von Texten für die meisten Benutzer sehr viel schneller möglich.
- Ein- und Ausgabehardware (Mikrofon, Lautsprecher) können sehr klein gebaut werden.
- Die Kommunikation kann berührungslos über größere Distanz erfolgen.
- Hände und Augen werden nicht beansprucht und können so für andere Aufgaben verwendet werden.

Kommerzielle sprachbasierte Lösungen finden sich bereits in einigen Spezialanwendungen. Einfache, sehr eingeschränkte und daher auch robuste *Kommandowort-erkenner* (Einzelworterkennung) helfen bei der Steuerung von Geräten. Gerade dort, wo Hände und Augen bereits anderweitig in Beschlag genommen sind, bietet sich diese Technik an. Ein gutes Beispiel hierfür ist die Bedienung des Autotelefons. Anstatt der mühsamen Eingabe der Telefonnummer reicht ein zuvor festgelegtes Wort für den zu erreichenden Teilnehmer und die Nummer wird herausgesucht und automatisch gewählt.

Kontinuierliche Spracherkenner sollen bei Textverarbeitungsprogrammen die alphanumerische Tastatur ersetzen (hörende Schreibmaschine). Der Wortschatz einer solchen Diktieranwendung muss entsprechend groß sein, um in akzeptabler Weise Briefe oder andere Schriftstücke erstellen zu können. Eine zusätzliche Schwierigkeit stellt der kontinuierliche Charakter der Eingabe dar. Die Wörter im Satz werden beim Sprechen oft nahtlos aneinander gefügt und daher müssen auch die Wortgrenzen bei der maschinellen Verarbeitung korrekt identifiziert werden. Im Vergleich zu Einzelworterkennern sind solche Systeme sehr aufwändig, benötigen viel Speicher und Rechenleistung. Der Nutzen ist momentan noch beschränkt, denn obwohl durch schnelle Algorithmen und Rechner die Erkennungsgeschwindigkeit inzwischen akzeptabel ist, sind die erreichten Fehlerraten immer noch zu hoch. Die Zeit, die anschließend zur Korrektur verwendet werden muss, übersteigt oft den Vorteil durch die zunächst beschleunigte Eingabe. Um überhaupt gute Erkennungsraten und hohe Geschwindigkeiten zu erreichen, werden diese Systeme nur in geräuscharmer Umgebung betrieben und in einer Einlernphase an einen bestimmten Sprecher angepasst.

Bei automatischen Auskunftssystemen möchte man einerseits einen Zugriff auf große Datenbestände oder einen großen Funktionsumfang ermöglichen, andererseits dafür nicht auf eine Tastatur angewiesen sein. Daher würde sich Sprache als Eingabemodalität auch hierfür

„ausgesprochen" eignen. Bei manchen Telefonvermittlungen oder Telefonauskunftssystemen findet man daher bereits heute Spracheingabe als Alternative zur Auswahl über das Tonwählverfahren. Intelligentere Systeme, die statt Ziffern oder vorgegebener Einzelwörter auch ausformulierte Anfragen verstehen, sind noch im Prototypenstadium. Anders als bei der Diktieranwendung muss nicht jedes Wort richtig verstanden werden, solange die Intention des Anrufers noch erkennbar ist. Die Schwierigkeiten liegen aber in der Qualität des Signals und der Interpretation der erkannten Wortfolge. Die Systemreaktionen aufgrund falsch erkannter oder unerwarteter Eingaben strapazieren oft genug die Geduld des Anrufers und werden daher mitunter etwas belächelt. Obwohl das kommerzielle Interesse an solchen Systemen sehr groß ist, ergeben sich im Vergleich mit menschlichen „Telefon-Operatoren" gewisse Akzeptanzprobleme bei den Benutzern.

Eine ähnliche Anwendung liegt im Falle von mobilen Informations- und Assistenzsystemen vor, für die in der hier vorgestellten Arbeit eine Spracherkennungskomponente entwickelt wurde. Darunter fallen zum Beispiel sprachbediente Autonavigationssysteme, für die größere Akzeptanzbereitschaft erwartet werden kann, da Alternativen in Form von menschlichen Dienstleistern fehlen. Auf diese Art Anwendung wird im folgenden Abschnitt eingegangen.

1.2 Sprachgesteuerte mobile Informationssysteme

Besonders unterwegs sind wir auf bestimmte Art von Information angewiesen. In einer neuen Umgebung müssen wir uns erst zurechtfinden. Als Tourist beispielsweise sucht man bestimmte sehenswerte Orte auf, möchte also wissen wie man dorthin gelangt und darüber hinaus geschichtliche Informationen erhalten. Als Geschäftsreisender sucht man zuweilen eine Unterkunft oder ein Restaurant. Wir müssen uns diese Informationen normalerweise aus den unterschiedlichsten Quellen zusammensuchen, was oft sehr umständlich ist. Ein kleines Beispiel soll dies verdeutlichen: Aus dem Reisehandbuch hat man ein Hotel ausgewählt und es nach längerem Suchen auf dem Stadtplan gefunden. Nachdem man sich lange erkundigt hat, wie man dort mit öffentlichen Verkehrsmitteln hingelangt, stellt man vor Ort fest, dass der Reiseführer veraltet ist und das Hotel ein Jahr zuvor geschlossen wurde.

Viele solcher Aufgaben könnten sich leichter mithilfe eines tragbaren, elektronischen Assistenten bewerkstelligen lassen. Er könnte unsere persönlichen Aufzeichnungen (wie Adressen, Kalenderdaten) sowie Informationen über die gerade besuchte Stadt speichern und abrufbar machen. Für weitergehende Anfragen (Hotel, Veranstaltungen) könnte das Gerät sich Informationen über Funk aus einer laufend aktuell gehaltenen Datenbank holen. Einige dieser Funktionen sind bereits in kommerziellen Geräten realisiert. Der Nutzen wird aber häufig noch durch die umständliche Bedienung gemindert. Große Tastaturen kommen kaum in Frage und so muss man sich oft mühsam durch umständliche Eingaberoutinen hangeln oder ist auf einen kleinen Funktionsumfang beschränkt. Aufgrund der oben genannten Eigenschaften wäre eine Bedienung durch Sprache äußerst vorteilhaft.

Ein ähnlicher Fall liegt vor, wenn wir im Auto unterwegs sind. Typischerweise sind wir hier auf der Suche nach dem besten Weg zu einem bestimmten Zielort. Wünschenswert wäre

auch zusätzliche und aktuelle Information über die Strecke (Baustellen, Umgehungen, Raststätten) und die Zielumgebung (Wetter, Unterkunft, Sehenswürdigkeiten). Hier ist die Beschränkung der Baugröße der Geräte nicht ganz so ausschlaggebend wie bei tragbaren Geräten. Hände und Augen sind aber bereits mit anderen Aufgaben beschäftigt, weshalb wiederum Sprachbedienung von großem Vorteil ist. In Zukunft könnten tragbares Gerät und Navigations- und Informationssystem im Auto auch verschmelzen oder sich zumindest abstimmen und ergänzen.

Warum werden aber bisher kaum sprachbasierte Lösungen angeboten? Den zuletzt vorgestellten Anwendungen ist gemein, dass sie mobil eingesetzt werden (auch wenn ein Autonavigationssystem dabei fest installiert ist), weshalb wir sie hier als ***mobile Informationssysteme*** bezeichnen wollen. Durch die beschränkte Größe und die meist sehr niedrig angesetzten Kosten für die Hardware kommen in naher Zukunft nur einfache Spracherkenner mit kleinem Speicher- und Rechenzeitbedarf in Frage. Es ist aber schon abzusehen, dass dies bald kein limitierender Faktor mehr sein wird. Größere Probleme liegen hingegen

1. in der Komplexität der Sprachverarbeitung und
2. in den im mobilen Umfeld auftretenden Störungen, die sich zudem auch noch ständig verändern.

Hauptthematik der vorliegenden Arbeit ist der zweite Punkt. Um trotz Störungen eine akzeptable Qualität der automatisch erkannten Spracheingaben zu gewährleisten, bedarf es einer ***robusten Erkennung***.

Im Laufe der hier vorgestellten Arbeit entstanden zwei Prototypen sprachbasierter mobiler Informationssysteme. Sie sollen als Basis dienen, diese Probleme anzugehen und um in einem realistischen Kontext Lösungen dafür zu finden. Im Rahmen des EU-Projektes VODIS[1] (*„Voice Operated Driver Information Systems"*, Sprachbediente Fahrerinformationssysteme) entstand ein spontansprachliches Navigationssystem für Autofahrer, das zunächst für die Demonstration in leiser Büroumgebung realisiert wurde. Durch die in dieser Arbeit vorgestellten Maßnahmen ist es nun prinzipiell auch möglich, Anfragen im fahrenden Auto mit ähnlicher Genauigkeit zu erkennen. Als zweiter Prototyp entstand aus der Navigationsanwendung ein tragbares Touristeninformationssystem, das neben Navigation auch Information über Sehenswürdigkeiten, Hilfe bei der Tourenplanung und sogar Übersetzung in andere Sprachen liefert. Ein ähnliches System wurde in Kooperation mit dem European Media Lab (Heidelberg) im Projekt *„TalkingMap"* (als Teil von *„DeepMap"*) realisiert, bei dem die gemachten Erfahrungen bezüglich des Spracherkenners und der Sprachverarbeitung zum Einsatz kamen.

1.3 Unterschiedliche Möglichkeiten der Spracheingabe

Das Auto hat im Laufe seines Bestehens eine gewisse Wandlung durchlebt: Das ursprünglich simple Transportmittel hat eine Reihe technischer Erweiterungen erfahren, die

[1] Projektnummer LE-2177

die Fortbewegung sicherer und bequemer gestalten (Beleuchtung, Scheibenwischer, Klimaanlage). Viele Menschen verbringen täglich viel Zeit im Auto, weshalb auch Unterhaltungs- (Radio, Kassettenspieler, CD-Spieler) und Kommunikationstechnik (Autotelefon) Einzug gehalten haben. Mittlerweile helfen erste Navigationssysteme bei der Routenfindung. Man findet also eine Fülle von Geräten und Funktionen, die meist manuell bedient werden. Oft ist der Fahrer durch das Übermaß an Knöpfen und Reglern überfordert, da er dazu genau die Sinne braucht (optische Kontrolle, Hände), die auch für seine Fahrsicherheit nötig sind.

Aus obigen Gründen bietet sich Spracheingabe als Alternative an, was im europäischen Projekt VODIS untersucht wurde. Der naheliegende Ansatz – von der Ingenieursseite betrachtet – ist die diversen Funktionen hierarchisch anzuordnen und durch bestimmte Befehlswörter auszusuchen [Pouteau und Arevalo 1998]. Das sieht für einige einfache Fälle so aus, dass man zunächst das gewünschte Gerät nennt, dort eine Funktion auswählt und, falls nötig, bestimmte Werte angibt. Die folgende Abbildung zeigt dafür zwei Beispiele gesprochener Benutzereingaben.

Abbildung 1.1: Gerätesteuerung mit dem „Prompted Approach"

Bereits hierfür ist mehr als ein einfacher Kommandoworterkenner mit festgelegtem Vokabular nötig. „Temperatur-20-Grad" könnte man noch als ein festes Kommandowort definieren, die Radiosender hingegen ändern sich von Region zu Region. Sie müssen also automatisch in den Wortschatz des Erkenners eingefügt werden. Dieses Vokabular ist jedoch zu jedem Zeitpunkt sehr eingeschränkt, was eine relativ zuverlässige Erkennung ermöglicht. Die Eingaben können direkt auf Funktionen der Geräte abgebildet werden.

Dabei ist der Benutzers in seiner Auswahl nicht frei, denn die möglichen Eingaben sind jederzeit durch das System bestimmt. Da es eine beträchtliche Anzahl von Gerätefunktionen geben kann, muss der Benutzer bei ihrer Auswahl unterstützt werden. Dies geschieht durch eine Liste, die durch das System ausgegeben wird. Daher spricht man von einem Ansatz mit **Systeminitiative** (in der englischen Literatur auch: „**Prompted Approach**"). Bei Telefonauskunftssystemen ist dieses Vorgehen ebenfalls üblich. Hier muss die jeweilige Liste dem Anrufer vorgelesen werden, was eine ziemlich langwierige Prozedur darstellt. Im VODIS-Projekt wurden die Auswahlpunkte daher auf einem kleinen Bildschirm auf der Fahrzeugkonsole angezeigt.

Dem gegenüber steht der Ansatz mit **gemeinsamer Initiative** von Benutzer und System („**Mixed Initiative Approach**"). Dem Benutzer werden dabei beliebige spontansprachliche Anfragen gestattet, die durch Rückfragen des Systems eventuell noch vervollständigt werden müssen. Dies entspricht eher der Art wie Menschen an eine Aufgabe heran gehen. Für das Beispiel der Temperatureinstellung würde man dann wie zu einem Beifahrer sagen:

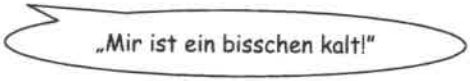


Abbildung 1.2: Gerätesteuerung durch spontansprachliche Anfrage[2].

Das erfordert vom System den Schluss, dass die Temperatur erhöht werden soll. Eventuell muss noch verhandelt werden, was mit „ein bisschen" gemeint ist.

Der Aufwand, den man für ein System betreiben muss, das diese Dinge beherrschen soll, mag für diese einfache Anwendung zu groß sein. Sobald die angebotenen Dienstleistungen komplexer werden, zum Beispiel bei der Eingabe von Fahrtzielen in das Navigationssystem, zeigt sich aber der Vorteil für den Benutzer deutlich. Die folgende Abbildung zeigt einen Vergleich der beiden Ansätze für diesen Fall. Für die Auswahl der Navigation ist beim „Prompted Approach" prototypisch eine Auswahlliste der vorgeschlagenen Eingabemöglichkeiten angegeben. Diese ändert sich nach jeder Eingabe und zeigt die neuen Möglichkeiten an.

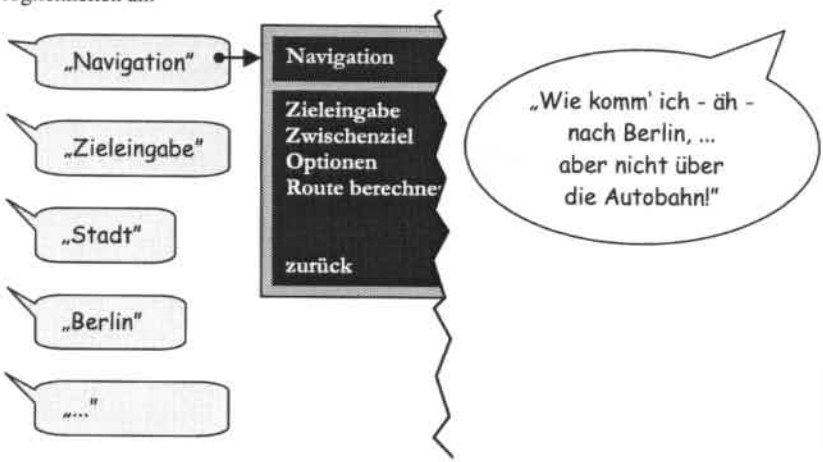


Abbildung 1.3: Vergleich von „Prompted Approach" (links) und „Mixed Initiative" (rechts) für eine Navigationsanfrage.

Der systemgesteuerte Ansatz führt zwangsläufig dazu, dass der Benutzer nach jedem Schritt auf den Bildschirm blickt und dadurch als Fahrer stark vom Verkehr abgelenkt wird. Mit der spontansprachlichen Anfrage ist man hingegen ohne große Anstrengung „gleich am Ziel".

Das Beispiel verdeutlicht allerdings auch wieder, wie schwierig die maschinelle Verarbeitung einer spontansprachlichen Anfrage sein kann. Kleine Versprecher und vielfältige Formulierungsmöglichkeiten sind typisch für spontansprachliche Anfragen. Sie erleichtern dem Sprecher die Aufgabe jedoch ungemein, da er sich weniger konzentrieren muss. Allerdings bringt „Mixed Initiative" mit freien spontansprachlichen Anfragen auch Nachteile mit sich. Abgesehen von seiner sehr schwierigen technischen Umsetzung ist es für den

[2] Hier eigentlich eine Feststellung.

Benutzer nicht einsichtig, welche Funktionen vom System tatsächlich angeboten werden. Es könnte ihm zum Beispiel entgehen, dass er auch die Möglichkeit hat, Zwischenziele zu bestimmen.

Aufgrund der technischen Schwierigkeiten wird sich in der Zukunft für mobile Informationssysteme wohl ein langsamer Wandel vom einfacher realisierbaren „Prompted Approach" hin zu „Mixed Initiative" vollziehen. Gerade im Auto liegt auch eine sehr schwierige akustische Umgebung vor, weshalb sich die Forschungsarbeiten auf diesem Gebiet auf die Verbesserung von Kommandoworterkennern[3] für den „Prompted Approach" oder die Ziffernerkennung[4] und Buchstabenerkennung[5] konzentrieren. Oft werden dazu künstlich verrauschte Testdaten verwendet. Systeme, die auch teilweise kontinuierliche Anfragen verwenden, sind selten (wie in [Muthusamy *et al.* 1999] zur Autonavigation, allerdings ohne Angabe von Erkennungsleistungen oder tatsächlicher Einsatzumgebung). In der vorliegenden Arbeit wurde ein Schritt weiter in Richtung zukünftiger Anwendungen getan und die Erkennung spontaner Sprache in der realen Autoumgebung untersucht. Techniken, die bereits erfolgreich zur Verbesserung von Einzelworterkennern in gestörter Umgebung eingesetzt wurden, erwiesen sich dabei als nur bedingt tauglich.

1.4 Störfaktoren im mobilen Umfeld

Mit welchen besonderen Schwierigkeiten hat ein Spracherkenner für spontansprachliche Anfragen im mobilen Umfeld zu kämpfen? Die Komplexität eines dafür benötigten kontinuierlichen Erkenners ist ähnlich der eines Diktiersystems. Zwar ist der benötigte Wortschatz meist wesentlich geringer, andererseits handelt es sich um spontane Sprache mit einer Reihe zusätzlicher Effekte wie Versprechern und umgangssprachlichen Verschleifungen. Durch Störungen aus der Umgebung wird der Sprecher eventuell abgelenkt und die Effekte verstärken sich. Die Struktur der Sprache ist ebenfalls weniger festgelegt („Schnell, zum Bahnhof möcht' ich!").

[3] [Song *et al.* 1998], [Fischer und Stahl 1998, 1999; Bippus *et al.* 1999], [Shozakai *et al.* 1998; Shozakai 1999], [Jabloun und Çetin 1999], [Gelin und Junqua 1999], [Ramalingam *et al.* 1999]

[4] [Schless und Class 1997, 1998], [Gong und Godfrey 1999], [Gelin und Junqua 1999]

[5] [Nguyen *et al.* 1999]

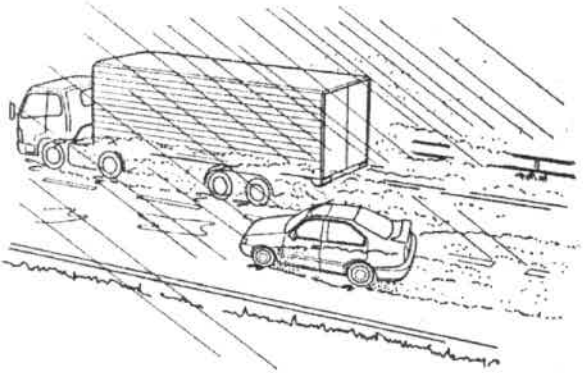


Abbildung 1.4: Schwierige und wechselnde Einsatzbedingungen für die automatische Spracherkennung im Auto.

Das **Hauptproblem** bei der Spracherkennung für mobile Informationssysteme stellen wechselnde Sprecher und Aufnahmeumgebungen dar. **Abbildung 1.4** macht deutlich, dass besonders im Auto unterschiedlichste Störungen auftreten können. Beim Überholen eines Lastwagens auf der Autobahnfahrt im Regen ergibt sich eine hohe Geräuschkulisse. Bei einer Rast abseits der Straße hingegen kann es völlig still sein. Die Störungen lassen sich grob in drei Kategorien einteilen, für die in der folgenden Auflistung einige Beispiele für die Autoumgebung aufgeführt sind:

1. Hintergrundgeräusche
   - bekannte Signale: Musik/Sprache aus Lautsprecher (Radio, Kassette, CD, Autotelefon)
   - vom Fahrzeug: Motor, Lüftung, Blinker, Scheibenwischer
   - vom Fahren: Wind, Straßenbelag
   - von außerhalb: Verkehr, Regen
2. Kanal
   - Raumakustik: Fahrzeugkabine, Position des Sprechers
   - Aufnahmehardware: Mikrofon
3. Sprecher
   - wechselnde Sprecher
   - Geräusche: Atmen, Räuspern
   - Sprechstil

Für das Training eines sprecherunabhängigen, kontinuierlichen Erkenners sind große Mengen an Sprachdaten nötig. Da eine solche Sammlung einen enormen Aufwand darstellt, versucht man, möglichst auf größere, bestehende Sprachdatenbanken zurückzugreifen. Diese enthalten meist mit guten Nahbesprechungsmikrofonen in geräuscharmer Laborumgebung aufgenommene Sprache. Die damit trainierten Erkenner sind für den Einsatz in einer solchen Umgebung optimiert. Ergeben sich im Einsatz andere Bedingungen (Schlagwort aus der englischen Literatur: „*Mismatch*"), stimmt das gelernte Modell nicht mehr mit der

angetroffenen Realität überein und schlechte Erkennungsleistungen sind die Folge. Die oben aufgeführten Störungen lassen sich dabei nicht sauber voneinander trennen und beeinflussen sich auch gegenseitig (**Abbildung 1.5**). So ist beispielsweise der Kanal im Sprachsignal nicht mehr von Eigenschaften des Sprechers zu trennen. Der Sprecher lässt sich auch vom Hintergrundrauschen beeinflussen, was zu einer veränderten Lautrealisierung führt (sogenannter Lombard-Effekt [Junqua 1996]). Daher ist es wichtig, bei Versuchen reale Daten zu verwenden, da sonst solche Wechselwirkungen nicht berücksichtigt werden können. Dennoch wird oft versucht, mit simulierten Daten (d.h. zum Beispiel künstlich zugemischtem Rauschen) die Verhältnisse auf einfache Art anzunähern. Dieser Weg wurde in dieser Arbeit nur in Ausnahmefällen beschritten, ansonsten wurde mit realen Daten gearbeitet.

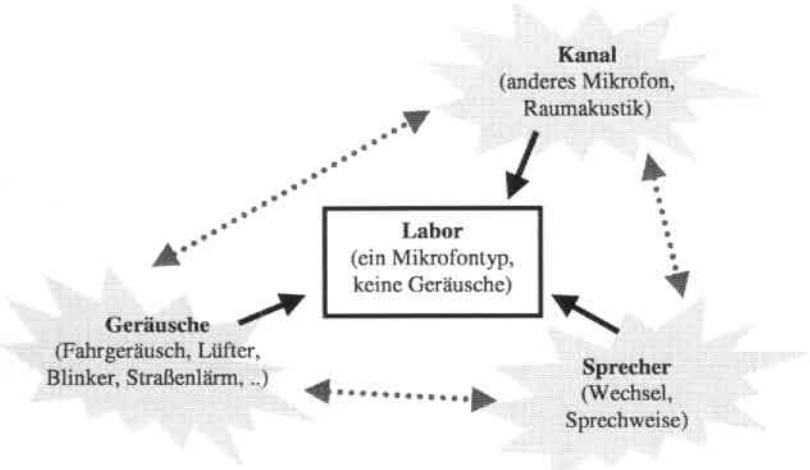


Abbildung 1.5: „Mismatch" beim Einsatz eines für Labordaten optimierten Erkenners in anderer Umgebung.

Da für die betrachtete Anwendung sprachgesteuerter mobiler Informationssysteme bisher keine geeigneten Sprachdatenbanken zur Verfügung stehen, wurde für die hier vorgestellten Experimente in einer umfangreichen Datensammlung entsprechendes Material gesammelt. Als wichtige exemplarische und äußerst schwierige Umgebung wurden spontansprachliche Anfragen in fahrenden Personenkraftwagen aufgenommen und verschriftet. Die unterschiedlichen Randbedingungen innerhalb dieses Szenarios (Fahrzeugtyp, Geschwindigkeit, Lüftereinstellung usw.) wurden nach Möglichkeit für die späteren Auswertungen protokolliert. Diese Daten wurden zum Testen verwendet, aber auch um Referenzsysteme zu trainieren. Letztere sollen einen Anhaltspunkt dafür geben, welche Erkennungsleistungen mit speziell gesammelten Daten und einem damit neu trainierten Erkenner erreichbar sind. Ziel dieser Arbeit ist es jedoch, basierend auf Labordaten die Robustheit gegenüber neuen und wechselnden Umgebungsbedingungen zu erhöhen, um so auch den Einsatz in anderen mobilen Anwendungen zu ermöglichen.

Als Vorgriff auf die Untersuchungen soll dem Leser veranschaulicht werden, wie drastisch die Leistungseinbußen eines bereits über mehrere Jahre optimierten Erkenners beim Einsatz in der Autoumgebung sein können. Der bisher in Büroumgebung eingesetzte, spontansprachliche Erkenner erreicht auf ungestörten Navigationsanfragen etwa 90 % Wortakkuratheit. Nach Drehen des Zündschlüssels (Lüftung, Motor) fällt das Ergebnis schon auf unter 80 %. Bei einer Geschwindigkeit von 100 km/h ist man bereits unter 40 %. Auch wenn in den spontansprachlichen Anfragen eine gewisse Redundanz steckt, reicht dieses Ergebnis nicht mehr aus, um die Benutzeranfragen genügend oft richtig zu interpretieren.

1.5 Zielsetzung und Aufbau der Arbeit

Ziel dieser Arbeit ist, basierend auf Trainingsdaten aus der Laborumgebung die Erkennungsleistung kontinuierlicher Spracherkenner für veränderte und wechselnde Umgebungsbedingungen zu verbessern. Nur so kann ein Einsatz der kontinuierlichen Spracherkennungstechnologie im mobilen Umfeld sinnvoll ermöglicht werden. Als typisches Szenario mit unterschiedlichen und sehr starken Störungen werden Anfragen an ein Navigationssystem im Auto betrachtet. Um sicherzustellen, dass der Erkenner sich auf neue und schnell wechselnde Situationen einstellen kann, werden bei den Erkennungsversuchen nur die Daten der jeweils aktuellen Äußerung benutzt. Vorangegangene Anfragen, Wissen über die Identität des Sprechers oder die Art der Umgebung sollen nicht verwendet werden. Der Erkenner muss sich dem Vergleich mit Systemen stellen, die sehr aufwändig mit speziell aus dieser Umgebung stammenden Sprachdaten trainiert wurden. Neben der Erkennungsrate sollen auch praktische Aspekte, wie die Erkennungsgeschwindigkeit bedingt durch den Rechenaufwand, oder Training und Optimierung eventueller Verfahren berücksichtigt werden.

Abbildung 1.6 zeigt eine Kapitelübersicht der vorliegenden Arbeit. Die Kapitel 2 bis 4 beschreiben die theoretischen und praktischen Grundlagen. Kapitel 2 geht dabei auf die allgemein übliche und auch hier benutzte HMM-Spracherkennungstechnologie ein und Kapitel 3 auf die später verwendeten Güte- und Optimierungskriterien. Die zum Teil selbst gesammelten Sprachdaten und ein Ausgangssystem zur Erkennung von Anfragen (Basissystem) werden in Kapitel 4 beschrieben.

Kapitel 5 bis 9 konzentrieren sich dann auf die *Robuste Spracherkennung*. Zunächst werden in Kapitel 5 die grundlegenden Ansätze vorgestellt und es erfolgt eine Einteilung der Verfahren. Einige Punkte werden dort in eigenen Abschnitten behandelt und zum Teil experimentell ausgewertet. Die Hauptbeiträge dieser Arbeit finden sich in den Kapiteln 6 bis 9 wieder. Sie behandeln die Anpassung der Erkennermodelle (Kapitel 6) und Kompensationsverfahren für wechselnde Sprecher (Kapitel 7) und den Übertragungskanal mit additivem Rauschen (Kapitel 8). Ausgehend von „State-of-the-Art"-Verfahren werden deren Schwachpunkte verbessert oder die Verfahren durch geeignetere neue Methoden ersetzt. In Kapitel 9 wird außerdem erstmalig ein Verfahren vorgestellt, das die Vorteile von Modellanpassung und Störkompensation miteinander kombiniert und mit dem alle

wesentlichen Störeinflüsse gemeinsam berücksichtigt werden können. Experimentelle Auswertungen belegen jeweils die Wirksamkeit für verschiedene Umgebungsbedingungen.

Das abschließende Kapitel 10 fasst die wichtigsten Erkenntnisse zusammen und gibt außerdem Hinweise darauf, wie man in einem Gesamtsystem, in dem ein Spracherkenner letztendlich zum Einsatz kommt, weitere Verbesserungen erzielen kann.

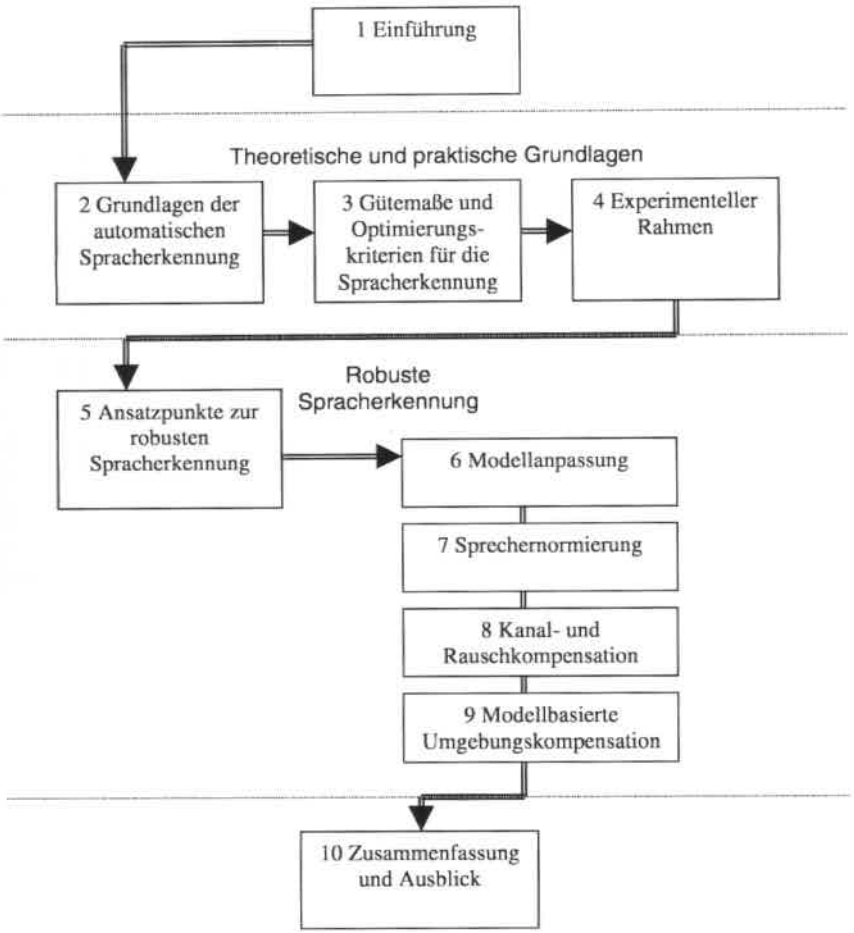


Abbildung 1.6: Kapitelübersicht

2 Grundlagen der automatischen Spracherkennung

„Am schwersten ist zu verstehen,
dass wir überhaupt etwas verstehen können."
– Albert Einstein -

Kapitelübersicht: In diesem Kapitel werden die Grundlagen der kontinuierlichen Spracherkennung erörtert, soweit sie für das Verständnis dieser Arbeit notwendig sind. Die Betonung liegt dabei auf der Merkmalsextraktion und der akustischen Modellierung. Diese Einführung beschränkt sich auf die Spracherkennung mithilfe sogenannter Hidden-Markov-Modelle, die sich gegenüber anderen Ansätzen, z.B. mit Neuronalen Netzen, durchgesetzt hat und heute Stand der Technik ist. Bezüglich weitergehender Literatur über automatische Spracherkennung sei hier auf einige deutschsprachige Werke [Holmes 1991, Boner 1992, Ruske 1994, Schukat-Talamazzini 1995] verwiesen. Zum Thema Sprache allgemein gibt [Crystal 1995] einen umfangreichen Überblick.

2.1 Grundprinzip der Spracherkennung

Die Spracherkennung ist das Gegenstück zur Spracherzeugung, bei der eine Folge von Wörtern $w = w_1, ..., w_W$ in ein akustisches Signal umgewandelt wird. Ziel der automatischen Spracherkennung ist es, aus dieser akustischen Repräsentation eine möglichst gute Hypothese w^* der zugrundeliegenden Wortfolge zu finden (siehe **Abbildung 2.1**).

Für die maschinelle Verarbeitung wird dafür zunächst das akustische Signal mit einem Mikrofon aufgenommen, digitalisiert und in eine Folge von Merkmalsvektoren $\mathbf{X}$ umgewandelt (*siehe Abschnitt 2.2 Merkmalsextraktion*).

Die geringste Fehlerrate ergibt sich nun bei der Wahl derjenigen Hypothese w^*, für die die bedingte Wahrscheinlichkeit $P(w|\mathbf{X})$ am größten ist (Maximum-a-posteriori-Prinzip). Mit der Bayes-Regel erhält man folgendes Entscheidungskriterium:

$$w^* = \arg\max_w \left(P(w|\mathbf{X})\right) = \arg\max_w \left(\frac{P(w)\cdot p(\mathbf{X}|w)}{p(\mathbf{X})}\right) = \arg\max_w \left(P(w)\cdot p(\mathbf{X}|w)\right) \qquad (2.1)$$

Mit einem Modell der Sprache (*siehe Abschnitt 2.3 Modellierung der Sprache*) werden die Wahrscheinlichkeit $P(w)$ der Wortfolge und die Wahrscheinlichkeitsdichte $p(\mathbf{X}|w)$ der Merkmalsvektorfolge, gegeben eine Wortfolge, näherungsweise bestimmt. Bei der Schätzung

der Wortfolgenwahrscheinlichkeit spricht man vom **Sprachmodell**[6], bei der Schätzung der Dichte p(**X**|*w*) vom **akustischen Modell** des Erkenners. Da zunächst die Modellparameter gelernt werden müssen, unterscheidet man bei einem automatischen Spracherkenner zwischen Trainings- und Erkennungsphase (*siehe Abschnitt 2.4 Training und Erkennung*). In der Trainingsphase werden bei gegebenen Wort- und Merkmalsvektorfolgen die Parameter des Modells gelernt. In der Erkennungsphase wird mit einem Suchalgorithmus anhand der Merkmalsvektorfolge und dem obigen Entscheidungskriterium eine Hypothese der gesprochenen Wortfolge gefunden.

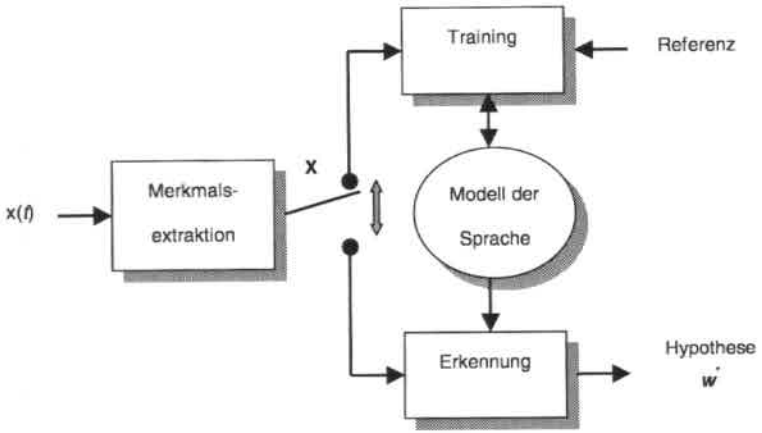


Abbildung 2.1: Übersicht der Spracherkennungskompenenten

2.2 Merkmalsextraktion

Dieser Abschnitt beschreibt die Grundlagen der Merkmalsextraktion, wie sie heutzutage für viele Laborerkenner verwendet wird. Obwohl immer wieder neue Verfahren vorgeschlagen und untersucht werden, hat sich über die letzten Jahre eine Hauptlinie herauskristallisiert, die auch mit dem *Janus Recognition Toolkit* (siehe Abschnitt 4.2) verfolgt wird. Das Grundkonzept wird auch hier hin und wieder variiert, um den verschiedenen Anwendungen, Sprachen und verfügbaren Datenmengen Rechnung zu tragen. Die folgenden

[6] Statt *Sprachmodell* werden zur genaueren Unterscheidung vom gesamten Erkennermodell der Sprache, was man ja ebenfalls als Sprachmodell bezeichnen kann, auch Begriffe wie *linguistisches Sprachmodell* verwendet (z.B. in [Schukat-Talamazzini 1995]). Die englische Literatur hat es in diesem Fall einfacher, da unterschiedliche Wörter für Sprache (engl: *language*) und gesprochene Sprache (engl: *speech*) existieren. Der hier klare englische Begriff „*language model*" wird aber meist als Sprachmodell geführt, was zwar missverstanden werden kann, aber dennoch üblich ist.

Ausführungen klammern die Erweiterungen zur Erhöhung der Robustheit aus, da sie als Hauptthema dieser Arbeit später noch gesondert behandelt werden.

Digitalisierung

Zunächst muss das akustische Signal zur Bearbeitung im Rechner digitalisiert werden. Die Schallwellen werden über ein Mikrofon aufgezeichnet, das elektrische Signal abgetastet und quantisiert. Bei der Abtastung ist darauf zu achten, dass die Abtastrate (gemessen in Hz) mindestens doppelt so groß ist wie die Bandbreite des Signals, da sonst Aliasing-Effekte auftreten. Daher wird das Eingangssignal eines Analog-Digital-Wandlers bandbegrenzt. Die Quantisierung (gemessen in bit) erfolgt entweder in gleichmäßigen Abständen (linear) oder mit einer Kompressionsfunktion, die größere Werte weniger fein quantisiert. Bei der Wahl der Abtastrate und Quantisierung orientiert man sich zunächst am Hörbereich und der Auflösungsfähigkeit des menschlichen Ohrs. Einerseits gibt es eine Obergrenze, bei der eine Erhöhung der Werte nicht mehr wahrgenommen wird, andererseits eine untere Grenze, bei der das Sprachsignal nicht mehr verständlich ist. Um eine effiziente Übertragung und Speicherung der Sprachsignale zu erzielen, hat man Interesse daran, die Abtastrate und die Quantisierungsgenauigkeit möglichst klein zu halten. Der Kompromiss reicht von 8 kHz Abtastrate für Telefonaufnahmen bis zu 44 kHz für HiFi-Aufnahmen, die neben Sprache auch Musik enthalten. Die Quantisierung liegt meist im Bereich zwischen 8 und 16 bit.

Sprachrahmen (Analyserahmen)

Um aus dem digitalen Zeitsignal $x(t)$ Merkmalsvektoren $\mathbf{x}[k]$ (also zeitlich aufeinander folgende Muster) zu gewinnen, werden mit einer Fensterfunktion $w(i)$ einzelne sogenannter Sprachrahmen k (engl.: *frames*) entnommen und mit einer Analysefunktion $f(\ldots)$ weiter verarbeitet.

$$\mathbf{x}[k] = f\left(w(0)\cdot x(k\cdot T),\ldots,w(\tau)\cdot x(k\cdot T+\tau),\ldots,w(L-1)\cdot x(k\cdot T+L-1)\right) \tag{2.2}$$

Diese Sprachrahmen (oder auch Analyserahmen, da genauso Pausensegmente verarbeitet werden) sollen einen aktuellen und gleichzeitig näherungsweise stationären Zustand des Sprachsignals repräsentieren. Dies gilt im Schnitt für Fensterbreiten von 16 bis 20 ms. Bei einer Abtastrate von $f_A = 16$ kHz und einer Fensterbreite von 16 ms ergeben sich zum Beispiel $L = 256$ Abtastwerte für jedes Fenster. Einfache Merkmale wie die mittlere Leistung und die Nulldurchgangsrate können mit einem Rechteckfenster berechnet werden. Bei einer Kurzzeit-Spektralanalyse werden hingegen zur Reduktion des Leck-Effektes zu den Rändern abfallende Fensterfunktionen wie das Hamming-Fenster verwendet (**Tabelle 2.1**, für weitere Fensterfunktionen siehe [Kammeyer und Kroschel 1989, Oppenheim und Schafer 1975]).

Um nun zu einer Folge von Merkmalsvektoren zu gelangen, werden die Fensterfunktionen meist in regelmäßigen Abständen über das Zeitsignal geschoben. Eine übliche Schrittweite T für die Analysefenster ist 10 ms, womit sich 100 überlappende Sprachrahmen und somit 100 Merkmalsvektoren pro Sekunde ergeben. **Tabelle 2.2** zeigt Beispiele vor allem spektraler Merkmale, die als Basis für weitere, darauf aufbauende Merkmale dienen. Der Schwellwert $c_{\min}$ in Gleichung (2.10) soll extreme Werte durch sehr kleine Leistungswerte $x_i^s[k]$ verhindern (die ja auch Null werden können).

Tabelle 2.1: In der Spracherkennung gebräuchliche Fenstertypen.

Fensterfunktion	$w(\tau) = \begin{cases} w^*(\tau) & \text{für } \tau = 0 \ldots L-1 \\ 0 & \text{sonst} \end{cases}$	
Rechteckfenster:	$w^*(\tau) = 1$	(2.3)
Hammingfenster:	$w^*(\tau) = 0{,}54 - 0{,}46 \cdot \cos\left(\frac{2\pi}{L-1} \cdot \tau\right)$	(2.4)

Tabelle 2.2: Beispiele von Spracherkennungs-Merkmalen mittels Kurzzeitanalyse.

Kurzzeit-Merkmale	$f(y(\tau))$ mit $y(\tau) = w(\tau) \cdot x(k \cdot T + \tau)$	
Mittlere Nulldurchgangsrate:	$x^{\text{cross}}[k] = \frac{1}{L-1} \sum_{\tau=1}^{L-1} \frac{1}{2} \cdot \left\lvert \text{sgn}(y(\tau)) - \text{sgn}(y(\tau-1)) \right\rvert$ mit $\text{sgn}(y(\tau)) = \begin{cases} 1 & y(\tau) \geq 0 \\ -1 & y(\tau) < 0 \end{cases}$	(2.5)
Mittlere Leistung:	$x^{\text{mpow}}[k] = \frac{1}{L} \sum_{\tau=0}^{L-1} y^2(\tau)$	(2.6)
Spektrum, diskrete Fouriertransformation (DFT):	$x_i^{\text{spec}}[k] = \sum_{\tau=0}^{L-1} y(\tau) \cdot \exp\left(-j \frac{2\pi}{L} \cdot \tau \cdot i\right)$	(2.7)
Betragsspektrum:	$x_i^{\text{m}}[k] = \left\lvert x_i^{\text{spec}}[k] \right\rvert$	(2.8)
Leistungsspektrum:	$x_i^{\text{s}}[k] = \left\lvert x_i^{\text{spec}}[k] \right\rvert^2$	(2.9)
Log-Spektrum:	$x_i^{\text{l}}[k] = \log\left(x_i^{\text{s}}[k] + c_{\min}\right)$	(2.10)

Geeignete Merkmale

Bei der Auswahl von geeigneten Merkmalen orientiert man sich an einem Modell der Spracherzeugung oder an der Funktionsweise des menschlichen Ohrs.

Mit der Spracherzeugung beschäftigt man sich schon sehr lange, insbesondere auch im Hinblick auf mögliche Syntheseverfahren. Je nach Art des Sprachlautes liegt eine Anregungsfunktion zugrunde, die mithilfe des Vokaltraktes geformt wird. So ergeben sich bei stimmhaften Lauten charakteristische Resonanzstellen im Spektrum des Signals, die sogenannten Formanten. Diese charakterisieren zwar den Laut recht gut, weisen jedoch

(beispielsweise durch anatomisch bedingte Unterschiede) eine gewisse Streuung auf. Die Anregung bei stimmhaften Lauten wird durch die Stimmbänder bewerkstelligt, ist aber für die Unterscheidung der Laute in den meisten Sprachen nicht relevant. Mit seinen starken Oberwellen dient das Anregungssignal nur als Träger, wobei die eigentliche Lautinformation als Gewichtung der spektralen Anteile übertragen wird. Da die mittlere Grundfrequenz dieser Anregung stark vom Sprecher abhängt (z.B. haben Frauen eine höhere Stimme), interessiert bei der Lauterkennung nur der grobe Verlauf des Spektrums.

Aus Untersuchungen der menschlichen Sprachrezeption hat man außerdem die folgenden Eigenschaften des Gehörs ermittelt:

- Die Frequenzauflösung, also die Fähigkeit zwei Töne ähnlicher Frequenz zu unterscheiden, nimmt mit steigender Frequenz nichtlinear ab.
- Die Lautstärkeempfindung nimmt mit steigender Signalenergie nichtlinear ab.
- Gegenüber Tönen steigender Frequenz nimmt die Empfindlichkeit oberhalb einer Grenzfrequenz nichtlinear ab.

Basierend auf diesen und anderen Überlegungen wurde eine Vielzahl von Methoden zur Gewinnung von Merkmalen entwickelt und untersucht. Immer wieder gibt es Veröffentlichungen, die mit einer modifizierten Merkmalsextraktion die Ergebnisse von Spracherkennern verbessern können. Die Wirksamkeit hängt aber von vielen Faktoren ab, beispielsweise der akustischen Modellierung oder der Art der Sprachdaten, und muss letztlich über die Fehlerrate oder ein anderes Kriterium (*siehe Kapitel 3*) für einen bestimmten Erkenner und ein definiertes Testset gefunden werden.

Weit verbreitet sind Merkmalsextraktionen basierend auf gehörsadäquaten Frequenzskalen (Einheit *Mel* oder Einheit *Bark*), die im Vergleich mit der linearen Frequenzskala in Hz die höheren Frequenzen immer gröber auflösen. **Abbildung 2.2** zeigt eine typische Näherungsfunktion. Der untere Frequenzbereich bis etwa 1000 Hz wird dabei etwa linear, der obere Frequenzbereich logarithmisch abgebildet. Um diese Art der Frequenzskalierung zu verwenden, werden zum Beispiel nach einer diskreten Fouriertransformation (Gleichung (2.7)) die Koeffizienten des Leistungsspektrums (Gleichung (2.9)) mittels einer *Mel*-Filterbank zusammengefasst. Die Mittenfrequenzen und Bandbreiten der einzelnen Bandpassfilter sind dabei in gleichmäßigen *Mel*-Abständen angeordnet. Dazu werden häufig dreieckige Filterformen verwendet, bei denen die Gewichtung der Koeffizienten von der unteren Grenzfrequenz linear bis zur Mittenfrequenz ansteigt und dann bis zur oberen Grenzfrequenz entsprechend wieder abfällt. Da es sich lediglich um eine gewichtete Zusammenfassung handelt, kann man das Ergebnis weiterhin als Leistungsspektrum betrachten. Die Ausgangswerte werden noch entsprechend Gleichung (2.10) logarithmiert, um auch die Lautstärkeempfindlichkeit anzupassen.

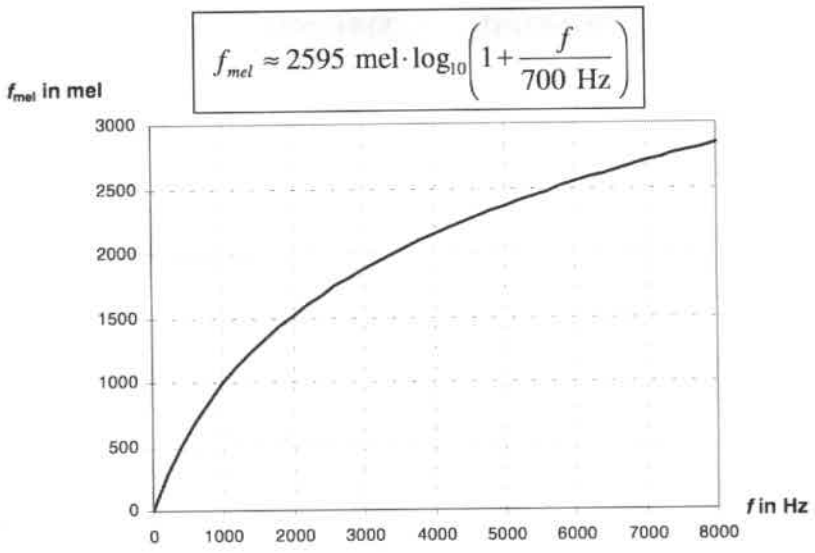


Abbildung 2.2: Umrechnung der linearen Frequenzskala in Hz in die gehörsadäquate *MEL*-Skala.

Auswahl der Merkmale

Neben den oben genannten Eigenschaften unterscheiden sich die Merkmale in folgenden Gesichtspunkten

- Diskriminanzfähigkeit
- Redundanz der Information
- Störanfälligkeit gegenüber Umgebungseinflüssen (Rauschen, Kanal, Sprecher)

Die ausgewählten Merkmale müssen eine gute Diskriminierung der (noch zu definierenden) akustischen Klassen ermöglichen. Um eine kompakte Parametrisierung des akustischen Modells zu erreichen, versucht man häufig auch die hohe Redundanz, die das Sprachsignal enthält, zu verringern. Je weniger Merkmale (d.h. Koeffizienten) der Merkmalsvektor enthält, umso schneller ist im Allgemeinen auch die Auswertung und damit der Erkennungsprozess. Besonders wichtig für eine robuste Erkennung ist es, die Abhängigkeit von Einflüssen zu reduzieren, die für die Erkennung der richtigen Wortfolge nicht wichtig sind. So ist zum Beispiel die stark durch das Geschlecht des Sprechers bestimmte Grundfrequenz der Anregungsfunktion ein gutes Merkmal für die Sprechererkennung, aber störend bei der Spracherkennung. Auch die Phase des Kurzzeitspektrums spielt keine Rolle, weshalb man sich auf das Betrags- oder Leistungsspektrum beschränkt.

Eine kompaktere Darstellung der Lautinformation ist mit dem sogenannten Cepstrum erreichbar. Da es sich hierbei um eine Frequenzanalyse des logarithmierten Spektrums handelt (mit der inversen diskreten Fouriertransformation), entstand der Name aus der teilweisen Umkehrung des englischen Wortes *„spectrum"*. Verwendet man als Ausgangspunkt das stets achsensymmetrische Leistungsspektrum, reduziert sich die Abbildung nach der Logarithmierung zu einer Multiplikation mit einer Cosinusmatrix $\mathbf{A}_{\cos}$ mit den Matrixelementen a_{ij} und der Dimension n.

$$\text{Cepstrum:} \quad \mathbf{x}^{c}[k] = \mathbf{A}_{\cos} \cdot \mathbf{x}^{l}[k] \qquad \text{mit} \quad a_{ij} = \cos\left(i \cdot (2j-1) \cdot \frac{\pi}{2n} \right) \tag{2.11}$$

Die ersten Koeffizienten des Cepstrums repräsentieren den groben Verlauf des Spektrums, weshalb man auch durch Nullsetzen der hinteren Koeffizienten (sogenanntes Liftering, von *„filtering"* stammend) die Anregungsfunktion bei stimmhaften Lauten ausblenden kann. Auch wenn diese Glättung schon teilweise durch die Anwendung der *Mel*-Filterbank geschehen ist, werden oft beide Verfahren kombiniert und man erhält die sogenannten *Mel*-Frequenz-Cepstralkoeffizienten (MFCC, von engl. *melfrequency cepstral coefficients*). Die Koeffizienten sind so wesentlich besser dekorreliert, was eine Verringerung ihrer Anzahl ermöglicht. Statt üblicherweise 16 bis 30 oder mehr *Mel*-Frequenz-Spektralkoeffizienten (MFSC) reichen 13 oder weniger MFCC aus, um ähnliche oder sogar bessere Erkennungsleistungen zu erzielen.

Kontextinformation

Da Spracherzeugung ein dynamischer Prozess ist, erhält der Erkenner durch die als statisch angenommene Information des aktuellen Sprachrahmens nur ein eingeschränktes Bild. Bestimmte Laute lassen sich leichter durch das Wissen über die Änderung aufeinander folgender Merkmalsvektoren identifizieren. Darum wird der Merkmalsvektor oft durch Näherungen der ersten und zweiten zeitlichen Ableitung der Grundmerkmale erweitert. Aus der Vektorfolge der $\mathbf{x}[k]$ werden dazu entweder Differenzenquotienten oder Ableitungen einer Polynomapproximation gebildet. Beide Formen werden in der Literatur als dynamische Merkmale [Furui 1986] oder auch als Delta-Koeffizienten $\Delta\mathbf{x}$ bezeichnet (üblicher Wert $d = 2$).

$$\text{Differenzenquotient:} \qquad \Delta\mathbf{x}[k] = \frac{\mathbf{x}[k+d] - \mathbf{x}[k-d]}{2d} \tag{2.12}$$

$$\text{Polynomapproximation:} \qquad \Delta\mathbf{x}[k] = \frac{\sum_{i=-d}^{d} i \cdot \mathbf{x}[k+i]}{\sum_{i=-d}^{d} i^{2}} \tag{2.13}$$

$$\text{Erweiterter Merkmalsvektor:} \qquad \mathbf{x}^{*}[k] = \begin{pmatrix} \mathbf{x}[k] \\ \Delta\mathbf{x}[k] \\ \Delta\Delta\mathbf{x}[k] \\ \cdots \end{pmatrix} \tag{2.14}$$

Neben den dynamischen Merkmalen werden oft noch weitere Merkmale wie Energie, Nulldurchgangsrate u.a. im erweiterten Merkmalsvektor aufgenommen.

Dimensionsreduktion

Durch die Erweiterung mit den dynamischen und anderen Merkmalen ist die Dimension des Merkmalsraumes angestiegen und es ist eventuell redundante Information enthalten. Unter Verwendung von Beispielen aus einer Trainingsmenge ließen sich nun durch reine Selektionsverfahren geeignete Koeffizienten auswählen. Besser ist jedoch ein Reduktionsverfahren, das den erweiterten Merkmalsvektor in einen Raum mit kleinerer Dimension abbildet. Ein bekannter Vertreter ist die Hauptachsentransformation, die zunächst die Koeffizienten durch eine lineare Abbildung dekorreliert. Anschließend können die Koeffizienten mit den größten Varianzen selektiert werden. Dieses Verfahren hat allerdings nur Sinn, wenn von vornherein gleichartige Merkmale benutzt werden. Ein Nachteil ist, dass die Diskriminanz, d.h. die Unterscheidbarkeit der Klassen, nicht berücksichtigt wird.

In der Spracherkennung hat sich eine andere lineare Abbildung als sehr viel nützlicher erwiesen. Sie wird durch die Lineare Diskriminanzanalyse (LDA) gefunden und optimiert ein Kriterium, das auf einer Klassenzuordnung der Trainingsbeispiele beruht (siehe Abschnitt 3.3). Da die LDA (wie auch andere Verfahren zur Dimensionsreduktion) von den verwendeten Trainingsdaten abhängt, spielen wechselnde Umgebungen eine wichtige Rolle. Die Auswirkung der LDA wird daher später noch genauer betrachtet (Abschnitt 5.4.2, S. 83 und Abschnitt 8.4.3, S. 144).

2.3 Modellierung der Sprache

2.3.1 Akustisches Modell

Die Aufgabe des akustischen Modells ist die Bestimmung der Wahrscheinlichkeitsdichte $p(\mathbf{X}|\boldsymbol{w})$. Bei Einzelworterkennern ist die Anzahl möglicher Wortfolgen $\boldsymbol{w}$ gerade die Größe des zu erkennenden Vokabulars (Die Wortfolge besteht immer nur aus einem Wort). Liegen genügend Trainingsbeispiele vor, ist es durchaus praktikabel, für jedes $\boldsymbol{w}$ (in diesem Fall jedes Wort) ein eigenes Modell zu halten und daraus eine Dichteschätzung zu berechnen. Bei der kontinuierlichen Spracherkennung existieren für die Wortfolge $\boldsymbol{w}$ sehr viele Möglichkeiten, so dass auf kleinere Einheiten (Wörter, Silben, Phoneme oder Subphoneme) zurückgegriffen werden muss, für die ausreichend viele Trainingsbeispiele vorliegen.

Die zeitliche Zuordnung der Untereinheiten zu einzelnen Merkmalsvektoren $\mathbf{x}[k]$ aus der Folge $\mathbf{X}$ und die Berechnung der Dichte für die gesamte Wortfolge $\boldsymbol{w}$ erfolgt dabei basierend auf der Theorie der *Hidden-Markov-Modelle* (HMM). Diese bestehen aus **Zuständen** z_s ($s = 1...S$) mit einer Anfangswahrscheinlichkeit π_s und einer Emissionswahrscheinlichkeit $b_{sk} = p(\mathbf{x}[k]|z_s)$ und **Übergängen** zwischen den Zuständen mit entsprechenden Übergangswahrscheinlichkeiten a_{ij}, dass vom Zustand z_i nach z_j gewechselt wird (siehe **Abbildung 2.3**). Die Eigenschaften eines HMMs lassen sich durch das Tripel $\lambda = (\boldsymbol{\pi}, \mathbf{A}, \mathbf{B})$ repräsentieren,

wobei die Parameter der Zustände und Übergänge dabei zu einem Vektor bzw. zu Matrizen zusammengefasst wurden.

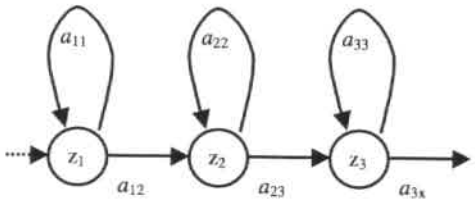


Abbildung 2.3: Einfaches Links-Rechts-HMM als Modell für eine sprachliche Einheit.

Für die Spracherkennung werden mögliche Wortfolgen durch kleinere, in ihrer Anzahl beschränkten Einheiten ersetzt. Diese Umwandlung erfolgt mithilfe eines Aussprachewörterbuches, in dem zum Beispiel für jedes Wort eine oder mehrere Phonemfolgen festgelegt sind. Phoneme sind die kleinste bedeutungsunterscheidende, aber doch nicht selbst bedeutungstragende lautsprachliche Einheit [Boner 1992]. Diese Einheiten werden nun durch einzelne oder mehrere HMM-Zustände repräsentiert, für die in der Trainingsphase die Parameter bestimmt wurden. Werden diese HMM-Bausteine immer nur sequentiell durchlaufen, handelt es sich also um Zustandsketten ohne rückwärts gerichtete Übergänge, spricht man von einem Links-Rechts-HMM. Um eine zeitliche Anpassung an die Folge der Merkmalsvektoren zu ermöglichen, sind jedoch Übergänge in denselben Zustand und das Überspringen von Zuständen in der Kette möglich. Setzt man diese Bausteine zusammen, erhält man ein HMM für komplexere Gebilde wie Wörter bis hin zu Wortfolgen. Hier werden dann auch häufiger parallel laufende Ketten verwendet (z.B. bei Aussprachevarianten).

Eine ausführliche Darstellung der Theorie der HMMs findet sich in [Rabiner 1989]. Für die weiteren Überlegungen der vorliegenden Arbeit ist vor allem die Modellierung der Emissionswahrscheinlichkeiten wichtig, weshalb nun darauf besonders eingegangen wird. Gesucht ist eine parametrische Näherung der bedingten Dichtefunktion $p(\mathbf{x}[k]|z_s)$, die der Emissionswahrscheinlichkeit b_{sk} entspricht. Um auch komplexere Dichten darstellen zu können, verwendet man häufig eine Mischverteilung basierend auf einfachen Dichtefunktionen, denn eine Dichte lässt sich als Summe von bedingten Einzeldichten folgendermaßen darstellen:

$$p(\boldsymbol{x} \mid z_s) = \sum_{m=1}^{M} P(m \mid z_s) \cdot p(\boldsymbol{x} \mid m, z_s) \qquad \text{mit} \quad \sum_{m=1}^{M} P(m \mid z_s) = 1 \tag{2.15}$$

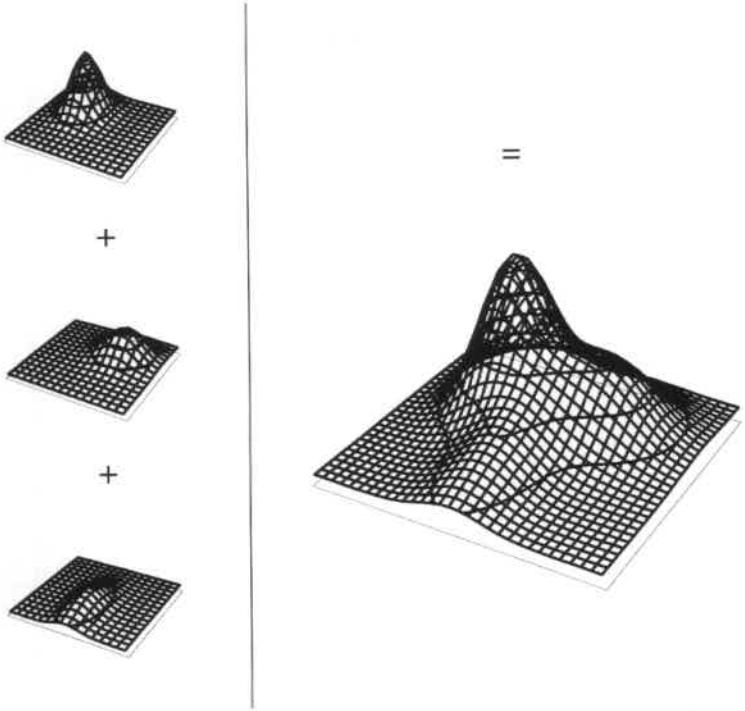


Abbildung 2.4: Beispiel einer 2-dimensionalen Mischverteilung mit 3 Gaußdichten.

Eine sehr gebräuchliche Näherung ist eine Mischverteilung aus multivariaten Gaußdichten (siehe auch Beispiel in **Abbildung 2.4**):

Mischverteilung:
$$\hat{p}(\boldsymbol{x} \mid z_s) = \sum_{m=1}^{M} \hat{P}(m \mid z_s) \cdot N(\boldsymbol{x}; \boldsymbol{\mu}_{s,m}, \mathbf{C}_{s,m}) \qquad (2.16)$$

Gaußdichte (Normalverteilung):
$$N(\boldsymbol{x}; \boldsymbol{\mu}, \mathbf{C}) = \frac{1}{\sqrt{(2\pi)^n \cdot |\mathbf{C}|}} \cdot e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \mathbf{C}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})} \qquad (2.17)$$

$$\text{mit} \quad \boldsymbol{\mu} = E\{\mathbf{x}\} \quad \text{und} \quad \mathbf{C} = E\left\{(\mathbf{x}-\boldsymbol{\mu}) \cdot (\mathbf{x}-\boldsymbol{\mu})^T\right\}$$

Die Parameter dieser Mischverteilung sind die Mixturgewichte $\hat{P}(m \mid z_s)$ sowie die Mittelwerte $\boldsymbol{\mu}_{s,m}$ und Kovarianzmatrizen $\mathbf{C}_{s,m}$ der Gaußdichten (die ebenfalls nur geschätzt werden, zur Vereinfachung aber ohne Dach geschrieben wurden). Sind die im Merkmalsvektor zusammengefassten n Merkmale nicht miteinander korreliert, erhält man eine diagonale Kovarianzmatrix $\mathbf{D}$. Die Dichte lässt sich für diesen Fall als Produkt der Einzelwahrscheinlichkeiten der unabhängigen Merkmale schreiben:

$$p(\boldsymbol{x} \mid z_s, m) = \prod_{i=1}^{n} p(x_i \mid s, m) \tag{2.18}$$

Die Gaußdichte kann man entsprechend als Produkt von eindimensionalen Gaußverteilungen ausdrücken:

$$N(\mathbf{x};\boldsymbol{\mu},\mathbf{D}) = \prod_{i=1}^{n} \frac{1}{\sqrt{(2\pi)\cdot\sigma_i^{\,2}}} \cdot e^{\frac{1}{2}\frac{(x_i-\mu_i)^2}{\sigma_i^{\,2}}} \tag{2.19}$$

mit der diagonalen Kovarianzmatrix:

$$\mathbf{D} = \begin{pmatrix} \sigma_1^{\,2} & 0 & \cdots & 0 & 0 \\ 0 & \ddots & & & 0 \\ \vdots & & \sigma_i^{\,2} & & \vdots \\ 0 & & & \ddots & 0 \\ 0 & 0 & \cdots & 0 & \sigma_n^{\,2} \end{pmatrix}$$

2.3.2 Sprachmodell

Ein statistisches Sprachmodell erlaubt die Schätzung der Wahrscheinlichkeit P(***w***) einer gegebenen Wortfolge ***w***. Da die zur Verfügung stehenden Ressourcen endlich sind und die meisten Anwendungen auch eine Beschränkung erlauben, kann für eine Erkennungsaufgabe ein **Vokabular** definieren werden. Dieses Vokabular legt die möglichen, in der Erkennung zu untersuchenden Wörter fest. Da für den hier betrachteten Ansatz auch eine Aussprache im Wörterbuch vorhanden sein muss, ist das Vokabular stets eine Untermenge der dort enthaltenen Einträge. Für die Einzelworterkennung verwendet man als Sprachmodell die A-priori-Wahrscheinlichkeit P(w) der Wörter w im Vokabular. Diese wird meist durch die relativen Häufigkeiten der Wörter in einer Trainingsmenge geschätzt.

Für eine Wortfolge kann die folgende Zerlegung durchgeführt werden[7]:

$$P(\boldsymbol{w}) = P(w_1,\ldots,w_W) = \prod_{i=1}^{W} P(w_i \mid \ldots, w_{i-2}, w_{i-1}) \tag{2.20}$$

Jeder Faktor dieses Produkts hängt von der bisherigen Vorgeschichte ab. Als gute Näherung haben sich *n-Gramme* bewährt, bei denen die verwendete Historie auf n - 1 Wörter beschränkt wird:

$$P(\boldsymbol{w}) \approx \prod_{i=1}^{W} P(w_i \mid w_{i-(n-1)},\ldots,w_{i-1}) = \prod_{i=1}^{W} \frac{P(w_{i-(n-1)},\ldots,w_{i-1},w_i)}{P(w_{i-(n-1)},\ldots,w_{i-1})} \tag{2.21}$$

Die einzelnen Faktoren können wiederum über die Häufigkeiten der entsprechenden Wortfolgen in großen Textkorpora oder aus den Transliterationen (Verschriftungen) einer Trainingsmenge geschätzt werden. Stehen ausreichende Mengen Textmaterial zur Verfügung, werden meist Trigramme ($n = 3$) benutzt. Für viele Wortfolgen werden dennoch keine oder

[7] Wörter mit negativem Index entfallen, um eine kompaktere Schreibweise zu ermöglichen.

nur wenige Beispiele vorhanden sein, weshalb man dann mit einer Rückfallstrategie (engl.: *back-off*) auf Bigramme ($n = 2$) oder Unigramme[8] ($n = 1$) ausweicht.

2.4 Training und Erkennung

Um das akustische Modell des Erkenners zu trainieren, müssen die beobachteten Muster $\mathbf{X} = \mathbf{x}[1]..\mathbf{x}[K]$ für eine gegebene Wortfolge w den einzelnen Klassen (Zuständen z_s) zugeordnet werden. Die Zustände entstammen einem HMM λ_w für die spezielle Wortfolge, das aus einzelnen HMM-Bausteinen zusammengesetzt wurde. Liegt die zeitliche **Zuordnung** $z[k]$ für jeden Analyserahmen k vor, können über einen Maximum-Likelihood-Ansatz die Parameter der Gaußdichten, die Mixturkoeffizienten und die Übergangswahrscheinlichkeiten bestimmt werden.

Bei der Erkennung entscheidet neben dem (linguistischen) Sprachmodell die akustische **Bewertung** $p(\mathbf{X}|\lambda_w)$ welche Wortfolge die wahrscheinlichste Hypothese darstellt. Dabei gibt $p(\mathbf{X}|\lambda_w)$ an, wie wahrscheinlich die Beobachtung $\mathbf{X}$ aus dem Modell für die Wortfolge generiert wurde und dient damit als Näherung für $p(\mathbf{X}|w)$.

Die Bewertung und die Zuordnung können mit dem *Forward-Backward-* und dem *Viterbi-*Algorithmus berechnet werden, die nun kurz vorgestellt werden.

Bewertung

Die Bewertung einer Musterfolge ergibt sich aus der Auswertung aller möglichen Zustandssequenzen $z = z[1]...z[K]$, wobei $z[k]$ jeweils einen Wert $s = 1...S$ abhängig vom Zustand z_s annehmen kann (siehe auch **Abbildung 2.5**):

$$p(\mathbf{X} \mid \lambda_w) = \sum_z p(\mathbf{X} \mid \lambda_w, z) = \sum_z \left(\pi_{z[1]} \cdot b_{z[1],1} \cdot \prod_{k=2}^{K} a_{z[k-1],z[k]} \cdot b_{z[k],k} \right) \tag{2.22}$$

Diese Art der Berechnung hat einen Aufwand $O(K \cdot S^K)$ bezüglich K, den man durch rekursive Berechnung der sogenannten Vorwärts-Wahrscheinlichkeit $\alpha_s[k]$ auf $O(K)$ reduzieren kann.

Vorwärts-Berechnung:

$$\alpha_s[k] = p(\mathbf{x}[1]\ldots\mathbf{x}[k], z[k] = s \mid \lambda_w) = \begin{cases} \pi_s \cdot b_{s,1} & \text{für } k = 1 \\ \sum_{i=1}^{S} \alpha_i[k-1] \cdot a_{i,s} \cdot b_{s,k} & \text{sonst} \end{cases} \tag{2.23}$$

$$p(\mathbf{X} \mid \lambda_w) = \sum_{s=1}^{S} \alpha_s[K]$$

[8] auch Monogramme genannt

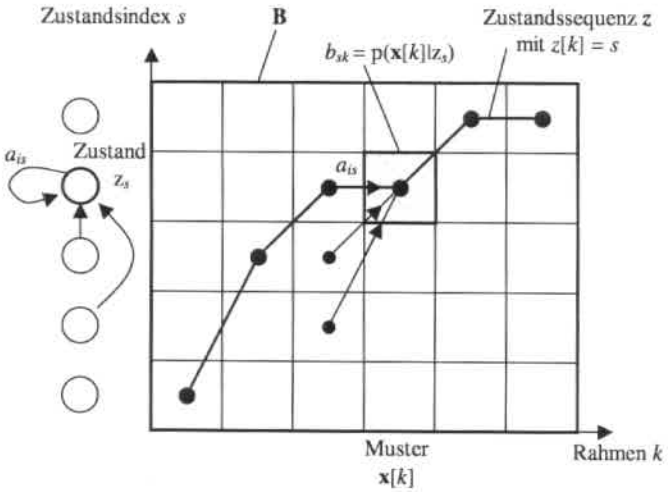


Abbildung 2.5: Zustandssequenz $\boldsymbol{z}$ als Pfad in der Matrix **B** eines HMMs λ_w.

Eine Näherung $p(\mathbf{X}, z^* \mid \lambda_w)$ für die Bewertung $p(\mathbf{X} \mid \lambda_w)$ erhält man über die wahrscheinlichste Zustandssequenz z^*, die mithilfe des Viterbi-Algorithmus gefunden werden kann.

$$z^* = \arg\max_z P(z \mid \mathbf{X}, \lambda_w) = \arg\max_z \frac{p(\mathbf{X}, z \mid \lambda_w)}{p(\mathbf{X} \mid \lambda_w)} = \arg\max_z \; p(\mathbf{X}, z \mid \lambda_w) \tag{2.24}$$

Viterbi-Algorithmus:

$$\begin{aligned} \alpha_s^{\vee}[k] &= \max_z \; p(\mathbf{x}[1] \ldots \mathbf{x}[k], z[k] = s \mid \lambda_w) \\ &= \begin{cases} \pi_s \cdot b_{s,1} & \text{für } k = 1 \\ \max_{i=1}^{S} \; \alpha_i^{\vee}[k-1] \cdot a_{i,s} \cdot b_{s,k} & \text{sonst} \end{cases} \end{aligned} \tag{2.25}$$

$$p(\mathbf{X}, z^* \mid \lambda_w) = \max_{s=1}^{S} \; \alpha_s^{\vee}[K]$$

Die praktische Umsetzung des Viterbi-Algorithmus geschieht meist ähnlich wie die der Vorwärts-Berechnung. Sie erfolgt ebenso zeitsynchron für alle Zustände vom Rahmen $k = 1$ bis zum letzten Rahmen $k = K$. Die Summation wird lediglich durch die Auswahl des Maximums ersetzt. Im Falle der hier vorliegenden HMMs für Sprache gelten noch einige Besonderheiten, die die Berechnung vereinfachen. So kommen nur Pfade (Zustandssequenzen) in Betracht, die im ersten Zustand beginnen und im letzten enden. Damit ist also nur $\pi_1 = 1$ und die Auswertung im letzten Schritt beschränkt sich auf die Vorwärtswahrscheinlichkeit des letzten Zustandes mit dem Index S. Außerdem sind viele

Übergangswahrscheinlichkeiten Null, wodurch immer nur einige wenige Vorgänger ausgewertet werden müssen (wie auch in **Abbildung 2.5** durch die Pfeile angedeutet).

Zuordnung

Um die wahrscheinlichste Zustandssequenz z^* zu erhalten, muss man sich bei jedem Schritt des Viterbi-Algorithmus für jeden Zustand den Übergang merken, der das maximale $\alpha_s^{\,v}[k]$ ergeben hat. Am Ende werden diese Übergänge zurückverfolgt, bis man den Ausgangszustand erreicht. Damit liegt eine eindeutige Zuordnung $z^*[k]$ jedes Musters zu einem Zustand vor, mit der man die entsprechenden Modellparameter neu bestimmen kann. Dazu definieren wir folgende Zuordnung für jeden Zustand:

Viterbi-Zuordnung:

$$\gamma_s^{\,v}[k] = \begin{cases} 1 & \text{für } z^*[k] = s \\ 0 & \text{sonst} \end{cases} \tag{2.26}$$

Kombiniert man die Vorwärts-Wahrscheinlichkeit $\alpha_s[k]$ mit der Rückwärts-Wahrscheinlichkeit $\beta_s[k]$, kann man die Wahrscheinlichkeit der Zustände zum Zeitpunkt k bestimmen. Man erhält hierdurch eine Zuordnung $\gamma_s[k]$, die der Wahrscheinlichkeitsverteilung über die Zustände zum Zeitpunkt k entspricht. Es müssen hierfür allerdings noch die Rückwärts-Berechnung und die kombinierte Berechnung (*Forward-Backward*) definiert werden:

Rückwärts-Berechnung:

$$\begin{aligned} \beta_s[k] &= \mathrm{p}(\mathbf{x}[k+1]\ldots\mathbf{x}[K] \mid z[k]=s, \lambda_w) \\ &= \begin{cases} 1 & \text{für } k = K \\ \sum_{i=1}^{S} a_{s,i} \cdot b_{i,k+1} \cdot \beta_i[k+1] & \text{sonst} \end{cases} \end{aligned} \tag{2.27}$$

$$\mathrm{p}(\mathbf{X} \mid \lambda_w) = \sum_{s=1}^{S} \pi_s \cdot b_{s,1} \cdot \beta_s[1]$$

Forward-Backward-Berechnung:

$$\mathrm{p}(\mathbf{X} \mid \lambda_w) = \sum_{s=1}^{S} \alpha_s[k] \cdot \beta_s[k] \quad \text{für beliebiges } k \tag{2.28}$$

Forward-Backward-Zuordnung:

$$\begin{aligned} \gamma_s[k] &= \mathrm{P}(z[k]=s \mid \mathbf{X}, \lambda_w) = \frac{\mathrm{P}(\mathbf{X}, z[k]=s \mid \lambda_w)}{\mathrm{p}(\mathbf{X} \mid \lambda_w)} \\ &= \frac{\alpha_s[k] \cdot \beta_s[k]}{\sum_{i=1}^{S} \alpha_i[k] \cdot \beta_i[k]} \end{aligned} \tag{2.29}$$

Zu jedem Zeitpunkt k gilt für beide Zuordnungsarten (*Forward-Backward* und *Viterbi*), dass die Summe der $\gamma_s[k]$ über alle Zustände 1 ergibt. Über die $\gamma_s[k]$ wird bestimmt, wie stark das Muster $\mathbf{x}[k]$ zur Berechnung der Parameter der Gaußmixturen eingeht (Ähnliches

gilt für die Neubestimmung der Übergangswahrscheinlichkeiten). Ausgehend von den $\gamma_s[k]$ muss noch eine Verteilung auf die einzelnen Gaußdichten m erfolgen, die durch $\gamma_{s,m}[k]$ festgelegt wird:

$$\gamma_{s,m}[k] = \mathrm{P}(z[k]=s,m \mid \mathbf{X},\lambda_w) = \frac{\mathrm{p}(\mathbf{x}[k],m \mid z_s,\lambda_w)}{\mathrm{p}(\mathbf{x}[k] \mid z_s,\lambda_w)} \frac{\mathrm{p}(\mathbf{X},z[k]=s \mid \lambda_w)}{\mathrm{p}(\mathbf{X} \mid \lambda_w)}$$
$$= \frac{\hat{\mathrm{P}}(m \mid z_s) \cdot \mathrm{N}(\boldsymbol{x};\boldsymbol{\mu}_{s,m},\mathbf{C}_{s,m})}{\sum_{n=1}^{M} \hat{\mathrm{P}}(n \mid z_s) \cdot \mathrm{N}(\boldsymbol{x};\boldsymbol{\mu}_{s,n},\mathbf{C}_{s,n})} \gamma_s[k] \qquad (2.30)$$

Training

Das Training des akustischen Modells erfolgt meist iterativ nach dem EM-Ansatz (*Expectation-Maximization* [Dempster *et al.* 1977], in diesem Zusammenhang auch Baum-Welch-Algorithmus genannt [Baum 1972]). Dabei wird abwechselnd die Zuordnung $\gamma_{s,m}[k]$ basierend auf dem bisherigen Modell λ_w ermittelt und dann werden über eine Maximum-Likelihood-Bestimmung (siehe Abschnitt 3.3.4) die Parameter eines neuen Modells λ'_w geschätzt. Die Statistiken, die sich, wie hier gezeigt, für das HMM λ_w einer Trainingsäußerung (Wortfolge $\boldsymbol{w}$) ergeben, werden allerdings für die komplette Trainingsmenge aufgesammelt und dann gemeinsam zur Neubestimmung der Einzelmodelle (HMM-Bausteine für Phoneme oder andere Einheiten) verwendet. Aus den jeweiligen Verschriftungen lassen sich die Satz-HMMs direkt aufbauen. Bei Einzelworterkennern bestehen die „Sätze" nur aus einzelnen oder zusammengezogenen Wörtern. Zu Beginn und am Ende werden jeweils noch Zustände für Sprachpausen eingefügt. Bei kontinuierlichen Erkennern kommen meist auch zwischen den einzelnen Wörtern noch optionale Pausenzustände hinzu, die auch übersprungen werden können.

Erkennung (Test)

Bei der Erkennung ist zunächst keine Wortfolge $\boldsymbol{w}$ wie beim Training gegeben, sondern sie soll gerade aus der Observation $\mathbf{X}$ und den trainierten Modellen bestimmt werden. Nur bei einer Auswertung des Erkenners (Test) wird die beste Hypothese $\boldsymbol{w}^*$ mit einer vorgegebenen Verschriftung verglichen. Um die wahrscheinlichste Hypothese nach Gleichung (2.1) zu finden, müssten theoretisch alle möglichen Wortfolgen bewertet werden. Im Falle von Einzelworterkennern sind dies gerade die Wörter im Vokabular und die Erkennung kann tatsächlich durch Auswertung aller Wortmodelle erfolgen. Bei der kontinuierlichen Spracherkennung kann am Ende eines jeden Wortes wieder ein neues beginnen. Die Wahrscheinlichkeit dafür kann über das Sprachmodell vorhergesagt werden. Statt eines Links-Rechts-HMM für eine bestimmte Wortfolge werden bei der kontinuierlichen Erkennung meist alle Wort-HMMs parallel verschaltet und es wird ein Übergang von den Wortenden in die Anfänge eingeführt (siehe **Abbildung 2.6**). Dessen Übergangswahrscheinlichkeit muss allerdings abhängig von der bisher betrachteten Wortfolge individuell bestimmt werden.

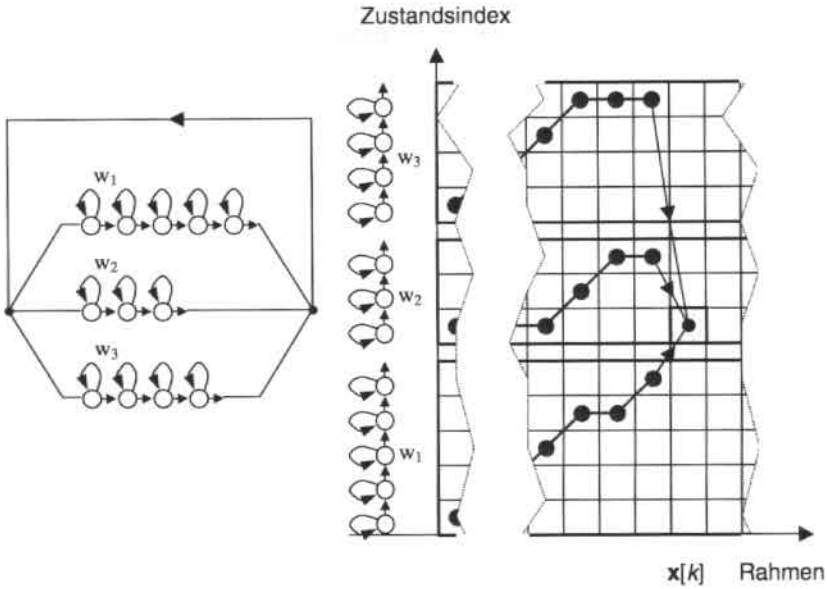


Abbildung 2.6: Verkettung der Wortmodelle durch Übergänge von den Wortenden zu den Wortanfängen.

Um die Zahl der zu untersuchenden Zustandssequenzen zu verringern, wird der Suchraum durch sogenannte *Pruning*-Techniken (von engl. *prune* = beschneiden) beschränkt. Dazu werden teilweise mehrere Suchdurchgänge verwendet und die Pfade (Zustandssequenzen bis zum aktuell untersuchten Sprachrahmen) mit zu schlechten Bewertungen von der weiteren Betrachtung ausgeschlossen. Hierauf wird konkreter bei der Behandlung des Basiserkenners in Abschnitt 4.2 eingegangen.

3 Gütemaße und Optimierungskriterien für die Spracherkennung

Das naive menschliche Denken geht von der Sache aus,
das wissenschaftliche von der Methode.
- Carl Friedrich von Weizsäcker -

Kapitelübersicht: In diesem Kapitel werden Maße definiert, mit denen die Leistungsfähigkeit eines Spracherkenners und seiner Teilkomponenten beurteilt oder optimiert werden können. Dabei wird nach der folgenden Definition zwischen Gütemaß und Optimierungskriterium unterschieden:

- **Gütemaß**: Dient zur Beurteilung der Leistung. Mit einem Gütemaß können unterschiedliche Erkenner oder Verfahren miteinander verglichen werden.
- **Optimierungskriterium**: Wird verwendet, um Parameter gezielt zu verändern. Wenn mehrere Optimierungskriterien gleichzeitig benutzt werden, müssen diese entweder unabhängig optimierbar sein oder gewichtet werden.

Zum Teil verschwimmen die Verwendungszwecke, denn von mehreren alternativen Verfahren wird man sich für dasjenige entscheiden, das bei einem bestimmten Gütemaß am besten abschneidet und optimiert auf längere Sicht das Spracherkennungssystem auf dieses Maß. Die ersten beiden Abschnitte (3.1 und 3.2) behandeln Gütemaße zur Beurteilung des gesamten Erkenner. Abschnitt 3.3 führt Maße auf, die sowohl als Beurteilung für Teilaspekte des Erkenners dienen können, aber auch bei diversen Verfahren gezielt zur Optimierung eingesetzt werden.

3.1 Erkennungsrate

Definitionen

Gebräuchlichstes Maß für die Bewertung eines Spracherkenners ist die Erkennungsrate bzw. ihr Pendant, die Fehlerrate. Für einen *Einzelworterkenner* ist die **Wortfehlerrate *WER*** (engl.: *word error rate*) das Verhältnis der Anzahl falsch erkannter Wörter N_{err} zur Gesamtzahl laufender Wörter N, die **Worterkennungsrate *WA*** (engl.: *word accuracy*) der verbleibende Teil:

$$WER = \frac{N_{err}}{N} \tag{3.1}$$

$$WA = 100\% - WER = \frac{N - N_{err}}{N} \tag{3.2}$$

Bei *kontinuierlichen Spracherkennern* ergeben sich neben verwechselten Wörtern (Substitutionen) auch Auslassungs- und Einfügefehler (engl.: *substitutions, deletions, insertions*). Zur Bestimmung der **Wortfehlerrate *WER*** werden all diese Fehler summiert und zur Anzahl der Referenzwörter *N* ins Verhältnis gesetzt.

$$WER = \frac{N_{err}}{N} = \frac{N_{sub} + N_{del} + N_{ins}}{N} \tag{3.3}$$

Bedingt durch die Einfügefehler kann die Anzahl der Fehler die Zahl der Referenzwörter übersteigen und es ergeben sich Wortfehlerraten über 100% (und eine negative *WA*). Die Anzahl der korrekten Wörter N_{cor} ergibt sich auch nicht mehr aus $N - N_{err}$, weshalb man zwischen **Worterkennungsrate oder Wortakkuratheit *WA*** (engl.: *word accuracy*) und **Wortkorrektrate *WC*** (engl.: *word correct*) unterscheidet:

$$WA = 100\% - WER = \frac{N - N_{sub} - N_{del} - N_{ins}}{N} \tag{3.4}$$

$$WC = \frac{N_{cor}}{N} = \frac{N - N_{sub} - N_{del}}{N} = WA + \frac{N_{ins}}{N} \tag{3.5}$$

Die Wortkorrektrate *WC* ist nicht unbedingt ein sinnvolles Maß, denn durch wiederholtes Auflisten des gesamten Erkennervokabulars (siehe Seite 26) kann man auch ohne tatsächliche Erkennung hohe Wortkorrektraten erzielen. Die Zahl der Einfügefehler ist dann zwar enorm (was bei *WC* keine Rolle spielt), aber das richtige Wort ist stets dabei (sofern überhaupt im Vokabular vorhanden).

Andererseits kommt es gerade bei verrauschten Daten oft zu mehreren Einfügefehlern oder Substitutionen hintereinander. Um eine höhere Wortakkuratheit zu erreichen, kann man diese Einfügungen durch geeignete Parameterwahl[9] stark reduzieren. Der Erkenner bevorzugt dann längere Wörter und macht statt mehreren kurzen Wörtern eher ein langes falsch. Dies geht aber auch auf Kosten der korrekten Wörter, was bei manchen Anwendungen (s.u., Abschnitt 3.2.2) störender ist als Einfügefehler. Daher sollte man auch die Wortkorrektrate neben der Wortakkuratheit im Auge behalten.

Beispiel zur Berechnung der Erkennungsraten:

Referenz:	berechne	den		Weg	zum	Bahnhof
Hypothese:	berechne	dem	ist	weg	zum	_____
Auswertung:	cor	sub	ins	sub	cor	del

Wir zählen 2 korrekte Wörter (cor) und 4 Fehler (sub, ins, del), bei 5 Referenzwörtern ergibt sich

[9] In den meisten kontinuierlichen Erkennern wird mit einer Wortübergangsstrafe die mittlere Wortlänge eingestellt. Dadurch können Auslassungs- und Einfügefehler gegeneinander reguliert werden.

die Wortfehlerrate $WER = \frac{4}{5} = 80\%$,

die Wortakkuratheit $WA = 1 - \frac{4}{5} = 20\%$ und

die Wortkorrektrate $WC = \frac{2}{5} = 40\%$.

Fehlerreduktion

Wird eine Modifikation an einem bestehenden Erkenner getestet, möchte man ein Maß dafür haben wie stark sich die Leistung verbessert hat. Absolute Gewinne sagen oft nicht viel aus, denn eine Verringerung der Fehler um absolut 1% bei einer kleinen Fehlerrate (sagen wir von 5% auf 4%) ist viel bedeutender als bei einer hohen Rate (von z.B. 51% auf 50%). Für solche Fälle definiert sich die **relative Fehlerreduktion** $ER_{relativ}$ (engl.: *error reduction*) einer Modifikation M als Verhältnis der Fehlerratendifferenz zur Fehlerrate des Vergleichssystems B (Basissystem):

$$ER_{relativ} = \frac{WER_B - WER_M}{WER_B} \tag{3.6}$$

Um Fehlerraten für die Ausgaben (Hypothesen) eines kontinuierlichen Spracherkenners zu berechnen, sind für die eingespeisten Äußerungen Referenztexte nötig. Diese bestehen wie die Hypothesen aus einer Verschriftung des Gesprochenen. Die Referenzen dienten jedoch entweder als Grundlage für die gesprochene (vorgelesene) Äußerung oder wurden durch Menschen nachträglich transliteriert. In beiden Fällen geht man von einer korrekten Umsetzung aus, was jedoch besonders bei spontansprachlichen oder verrauschten Äußerungen nicht immer zutrifft. Eine offene Frage ist hier auch, ob z.B. Häsitationen (wie „äh") oder abgebrochene Wörter verschriftet werden. Zum Teil ergeben sich auch Probleme bei der Definition der Einheiten: In manchen Sprachen kennt man keine Worteinheiten wie im Deutschen und selbst hier stellt sich die Frage, wie man mit langen, zusammengesetzten Wörtern umgehen soll.

OOV-Rate

Die für einen kontinuierlichen Spracherkenner theoretisch erreichbare Untergrenze der Wortfehlerrate wird durch die sogenannte *OOV*-Rate (engl.: *out of vocabulary rate*) bestimmt. Selbst bei noch so großem Vokabular kann es für bestimmte Anwendungen (spontansprachliche Navigationsanfragen sind ein Beispiel dafür) vorkommen, dass Begriffe verwendet werden, die nicht im Vokabular enthalten sind. Die *OOV*-Rate wird für ein gegebenes Testset bestimmt. Es ist das Verhältnis der nicht im Vokabular vorkommenden, laufenden Wörter N_{unk} (von engl.: *unknown*) zur Gesamtzahl N der Wörter im Testset. Wenn also ein unbekanntes Wort mehrfach vorkommt, geht es auch mehrfach in die *OOV*-Berechnung ein. Da der Erkenner keine Chance hat, diese Wörter richtig zu hypothetisieren, ist die *OOV*-Rate die minimal erreichbare Fehlerrate.

$$OOV\text{-}Rate = \frac{N_{unk}}{N} \tag{3.7}$$

Meist ist ein *OOV*-Wort sogar für mehrere Fehler verantwortlich. Werden keine besonderen Maßnahmen getroffen, so werden an der betreffenden Stelle ein, oft sogar mehrere kurze, einigermaßen passende Ersatzwörter erkannt. Durch die Bindung des Sprachmodells können auch bei den folgenden Wörtern weitere Fehler verursacht werden. Mit einem Konfidenzmaß kann man versuchen falsch erkannte Wörter zu detektieren, um so wenigstens die Folgefehler zu reduzieren. Eine andere Möglichkeit stellen sogenannte *„mumble words"* (Gemurmel) dar. Sie sind mit einem allgemein gehaltenen akustischen Modell repräsentiert und werden somit erkannt, wenn keines der spezielleren Wortmodelle besser passt.

3.2 Anwendungsspezifisches Gütemaß

3.2.1 Echtzeitfaktor und Speicherbedarf

Neben der Güte der Ausgaben eines Erkenners, die mit der Erkennungsrate gemessen werden kann, spielen für praktische Anwendungen noch andere Kenngrößen eine Rolle. Gerade für mobile Anwendungen ist die Hardware heute noch ein beschränkender Faktor. Durch das größere Vokabular und den aufwändigeren Erkennungsprozess verbrauchen kontinuierliche Erkenner deutlich mehr **Arbeitsspeicher** als einfache Kommandowörterkenner. Für die Benutzbarkeit ist vor allem die Geschwindigkeit des Erkenners entscheidend. Als Maß dafür verwendet man den **Echtzeitfaktor *RT*** (engl.: *real-time factor*). Er ist definiert als das Verhältnis der Verarbeitungszeit zur Dauer der zu erkennenden Äußerung. Bei der Zeitdauer für die Erkennung kann man noch unterscheiden zwischen reiner Rechenzeit (wie lange hat der Prozessor mit der eigentlichen Erkennungsaufgabe verbracht) und der tatsächlichen Zeit, die benötigt wurde. Letztere ist meist länger, besonders wenn noch andere Programme auf dem Prozessor laufen oder aufgrund knappen Speichers Daten oft aus- und wieder eingelagert werden müssen. Da die Prozessorrechenzeit konsistentere Werte liefert, wird meistens sie als Verarbeitungszeit verwendet. Der Anteil des benutzten Speichers und besonders der Echtzeitfaktor sind stark rechnerabhängig. Daher werden sich die hier noch bestehenden Probleme im Laufe der Zeit durch schnellere Rechner mit mehr Speicher verringern und spielen für Forschungsarbeiten eine etwas untergeordnete Rolle. Um jedoch die automatische Spracherkennung für reale Anwendungen zu untersuchen und um Versuchsreihen schnell umzusetzten, sind diese Faktoren dennoch nicht ganz außer Acht zu lassen.

Prinzipiell sind kleine Echtzeitfaktoren (möglichst kleiner als 1) zu bevorzugen. Diese werden aber häufig nur auf Kosten von Erkennungsrate und erhöhtem Speicherbedarf erreicht. Je nach Anwendung muss man also die unterschiedlichen Anforderungen an die Echtzeitfaktoren der Erkenner mit den anderen Werten abwägen. Beim *Online*-Diktat (hörende Schreibmaschine) sind die Anforderungen diesbezüglich hoch, denn die Erkennung

soll schritthaltend und der Echtzeitfaktor kleiner als 1 sein. **Schritthaltend** bedeutet, dass noch während des Sprechens mit der Erkennung der Äußerung begonnen wird. Somit kann das System kontinuierlich Ausgaben liefern und der Sprecher muss zwischendurch nicht warten. Bei einer *Offline*-Anwendung (z.B. *Offline*-Diktat oder automatische Indexierung einer Videodatenbank [Kemp *et al.* 1998]) kommt es meist weniger auf die Geschwindigkeit an. Hier wartet der Sprecher nicht auf unmittelbare Ausgaben des Spracherkenners. Bei einem sprachbasierten Informationssystem sind kleinere Wartezeiten tolerierbar. Der Echtzeitfaktor sollte jedoch nahe genug bei 1 liegen, um die Geduld des Benutzers nicht allzu sehr zu strapazieren. Eine schritthaltende Verarbeitung wäre auch hier wünschenswert, jedoch braucht der Erkenner auch genügend Sprachmaterial, um sich auf wechselnde Sprecher und Umgebungsverhältnisse einzustellen.

3.2.2 Gesamtperformanz

Bei einem Diktiersystem entscheiden über den Nutzen einer Spracheingabe die Geschwindigkeit und die Anzahl der noch zu korrigierenden Fehler. Mit dem Echtzeitfaktor und der Wortfehlerrate sind also schon zwei Gütemaße erwähnt, die diese Eigenschaften beschreiben.

Für Informationssysteme ist es nicht nötig alle Wörter einer Anfrage richtig zu erkennen, solange die Intention des Benutzers noch aus der Hypothese des Erkenners ersichtlich ist. Die häufig auftretenden „den"/„dem"-Verwechslungen (ebenso „zur"/„zum") sind für das Verständnis nicht entscheidend, wohl aber zum Beispiel der Name einer gesuchten Straße. Einfügefehler am Anfang oder Ende einer durch Rauschen gestörten Anfrage sind eventuell nicht weiter störend, da durch die Analyse des Satzinhaltes diese nicht in Bezug mit den restlichen Wörtern gebracht werden können und somit bei der Interpretation ignoriert werden. Letztlich ist entscheidend, ob das Informationssystem richtig auf die Anfrage reagiert. Eine solche Auswertung ist allerdings sehr aufwändig und hängt stark von den verwendeten Systemkomponenten ab. Auch eine unterschiedliche Gewichtung von Wortfehlern wurde in der Vergangenheit als Alternative zur einfachen Wortfehlerrate in Betracht gezogen, ist aber genauso abhängig von einer bestimmten Anwendung und den verwendeten Systemkomponenten.

Da man davon ausgehen kann, die Gesamtperformanz des Systems durch eine Erhöhung der Wortakkuratheit zu verbessern, bleibt letztere trotz der aufgeführten Punkte eine der objektivsten und universellsten Möglichkeiten zur Optimierung des Spracherkenners innerhalb eines Gesamtsystems.

3.3 Andere Maße

Im Gegensatz zu den Gütemaßen der vorigen Abschnitte für den gesamten Erkenner sollen nun Maße behandelt werden, die dazu dienen, im Erkenner benutzte Verfahren zur Robustheitssteigerung zu bewerten. Sie werden verwendet, um diese Verfahren zu vergleichen oder ihre noch unbestimmten Parameter einzustellen. Zum Teil werden sie auch als Grundlage der Verfahren selbst benutzt, um durch die Optimierung des Maßes automatisch

Parameter zu bestimmen. Diese drei Verwendungszwecke sollen zunächst noch einmal etwas ausführlicher beleuchtet werden. Die aufgeführten Beispiele für Verfahren werden zum Teil in den Kapiteln 5 bis 6 eingehender behandelt und die Maße in den folgenden Abschnitten definiert.

- Der beste *Vergleich von Verfahren* ergibt sich meist aus der jeweiligen Erkennungsrate auf einem bisher nicht verwendeten Testset (d.h. die Daten wurden nicht zum Trainieren oder Optimieren benutzt). Da die Erkennungsraten das Ergebnis komplexer Prozesse sind, ist es oft hilfreich Maße zu betrachten, die direkter mit den zu verbessernden Teilkomponenten zusammenhängen. Ein Beispiel hierfür sind Geräuschreduktionsfilter (Abschnitt 8.4, S. 137), deren Wirkungsweise auf die Cepstrale Distanz, ein quadratisches Fehlermaß für cepstrale Merkmale, verglichen werden kann (S. 145).
- Viele Verfahren besitzen freie, noch *zu optimierende Parameter*, die generell oder für spezielle Verwendungszwecke („von Hand") bestimmt werden müssen. Das kommt zum Teil daher, dass sie auf einem vereinfachten Modell der Sprache oder der Umgebung aufbauen und daher durch ausgleichende Maßnahmen korrigiert werden müssen. Dazu werden weiter unten Optimierungskreisläufe besprochen. Beispiele hierfür sind wiederum Geräuschreduktionsfilter, die sich oft durch eine Fülle von noch zu bestimmenden Parametern auszeichnen. Die freien Parameter werden meist mithilfe des Signal-zu-Rauschverhältnisses eingestellt.
- Definiert man ein Maß, das für eine Teilkomponente des Erkenners berechnet werden kann und bei dem man davon ausgeht, dass es mit der Gesamtperformanz korreliert, kommt es als *Optimierungskriterium* für ein automatisches Verfahren in Frage. Mittels einer geeigneten Modellvorstellung setzt man dann zum Beispiel eine bestimmte Art von Abbildung voraus und bestimmt deren Parameter über das Kriterium. Besonders für lineare Abbildungen ergeben sich häufig geschlossene Lösungen. Beispiele hierfür sind die LDA (Lineare Diskriminanzanalyse, S. 81) zur Merkmalsextraktion oder die MLLR (*Maximum Likelihood Linear Regression*, S. 96) zur Adaption des akustischen Modells. Auch für Ballungsverfahren werden Optimierungskriterien verwendet. So zum Beispiel wird beim Basic-Isodata (oder *k-Means*) der mittlere quadratische Fehler zu den Ballungszentren minimiert. Bei der Ballung von phonetischen Klassen wird häufig ein Entropie-Maß verwendet.

Optimierungskreisläufe zur Parameteroptimierung

In Abschnitt 3.1 wurde die **Erkennungsrate**, das wohl wichtigste Maß zur Bewertung eines Spracherkenners, behandelt. Um die Parameter eines Verfahrens zur Verbesserung der Spracherkennung einzustellen, kann man in einigen Fällen direkt die Erkennungsrate verwenden. Dies führt im Rahmen der Möglichkeiten des Verfahrens zum optimalen Ergebnis. Da normalerweise kein einfacher funktionaler Zusammenhang zwischen Erkennungsrate und Parameter gegeben ist, müssen, gesteuert durch ein Optimierungsverfahren, auf einer Kreuzvaldierungsstichprobe die Erkennungsraten für die untersuchten Parameter bestimmt werden.

Im einfachsten Fall liegt hierfür ein Erkenner vor, der mit ungestörten Daten trainiert wurde und nun mit gestörten Daten getestet wird. Beseitigt das Verfahren die Störungen, indem es versucht die ungestörten Merkmale wieder herzustellen, reicht eventuell für jede Parametereinstellung ein Erkennungslauf aus. Verändert das Verfahren (bzw. seine Parameter) die Merkmale von Grund auf, muss zuvor ein Training der akustischen Modelle des Erkenners durchgeführt werden. Liegen Daten zum Trainieren aus der gestörten Umgebung vor, empfiehlt sich dies für fast alle Verfahren, denn so werden auch die Reststörungen (die es praktisch immer gibt) in den Modellen berücksichtigt. Dieses Vorgehen führt letztlich zu optimalen Erkennungsraten für diese spezielle Umgebung, ist aber sehr aufwändig, denn:

- *Dieser „**große Optimierungskreislauf**" (siehe **Abbildung 3.1**) schließt für jede zu prüfende Parametereinstellung ein vollständiges Neutraining des Erkenners ein (bei kontinuierlichen Erkennern zum Teil mehrere Tage Rechenzeit) und einen meist ebenfalls aufwändigen Testlauf. Mit einem Optimierungsverfahren wird die Wortfehlerrate für ein Kreuzvalidierungs-Testset minimiert.*

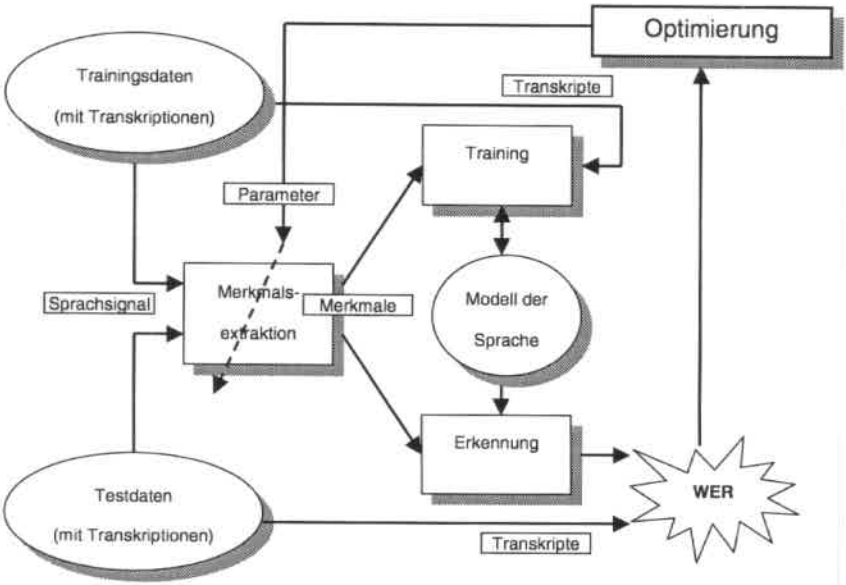


Abbildung 3.1: Großer Optimierungskreis zur Parametereinstellung eines Verfahrens in der Merkmalsextraktion über die Wortfehlerrate WER.

Verwendet man hingegen ein alternatives Maß, das sich anhand einer Stichprobe für einzelne Teilkomponenten des Erkenners einfach berechnen lässt, reduziert sich der Aufwand (besonders bei komplexeren Systemen) beträchtlich:

- *Dieser „**kleine Optimierungskreislauf**" (siehe **Abbildung 3.2**) erfordert für jede zu prüfende Parametereinstellung die Berechnung eines Gütemaßes. Mit einem Optimierungsverfahren wird dieses Maß auf einem Kreuzvalidierungs-Testset optimiert.*

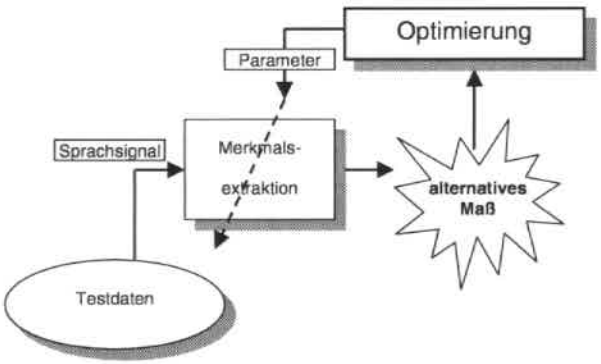


Abbildung 3.2: Kleiner Optimierungskreis zur Parametereinstellung eines Verfahrens in der Merkmalsextraktion über ein alternatives Maß.

Verwendet man einen kleinen Optimierungskreislauf, ist nicht garantiert, dass sich damit auch gute Erkennungsraten erzielen lassen. In den meisten Fällen kann aber so zumindest eine gute Voreinstellung der Parameter erreicht werden, die man dann nochmals anhand der Erkennungsrate bewerten kann. Die obigen Abbildungen zeigen als Beispiel die Optimierung eines Parameters in der Merkmalsextraktion, da letztere für die vorliegende Arbeit besonders relevant ist. Der kleine Optimierungskreis wird aber häufig auch für andere Teile des Erkenners benutzt. Mit der *Perplexität* einer Kreuzvalidierungsstichprobe (nur Verschriftung) lassen sich zum Beispiel statistische Sprachmodelle einfach und schnell bewerten. Gerade für dieses Beispiel ist allerdings die Übertragbarkeit auf die Erkennungsrate oft nicht gegeben.

In den nächsten Abschnitten werden folgende für die vorliegende Arbeit relevanten Maße behandelt:

- Signal-zu-Rausch-Verhältnis (SNR)
- Mittlerer Quadratischer Fehler (MSE)
- Lineare Diskriminanz (LD)
- Likelihood

3.3.1 Signal-zu-Rausch-Verhältnis (SNR)

Kurzbeschreibung: *Das Signal-zu-Rausch-Verhältnis beurteilt, wie stark ein Signal verrauscht ist. Je nach Berechnungsvorschrift muss hierfür auch das ungestörte Signal oder das reine Rauschsignal zur Verfügung stehen. Je größer der SNR-Wert, umso höher die Qualität des Signals.*

Definition

Gegeben sei ein digitalisiertes Zeitsignal $x(t)$. Dieses setzt sich additiv aus der Nutzsignalkomponente $s(t)$, folgend auch kurz Signal genannt, und einem Störsignal $n(t)$ zusammen. Es gilt also:

$$x(t) = s(t) + n(t) \tag{3.8}$$

Das Signal-zu-Rauschverhältnisses (SNR von engl. *Signal-to-Noise-Ratio*) ist definiert als der Quotient aus Signalvarianz σ^2_s und Störvarianz σ^2_n. Die SNR-Werte werden meist in Dezibel (dB) angegeben, was zur folgenden Definition führt:

$$SNR = 10 \cdot \log_{10} \frac{\sigma^2_s}{\sigma^2_n} \text{dB} \tag{3.9}$$

SNR für Sprachsignale

Da wir es hier mit mittelwertfreien[10] Zeitsignalen zu tun haben, kann statt der Varianz auch die Energie oder die Leistung verwendet werden. Mit dieser Definition ergeben sich allerdings bei Sprachsignalen zwei Probleme, weshalb einige Erweiterungen nötig sind:

- Das Maß nach Gleichung (3.9) ist eine gute Beschreibung für die Signalqualität im Falle von stationären Signalen. Bei Sprache handelt es sich allerdings nicht um solche, weshalb man quasistationäre Bereiche betrachtet. Aus dem Zeitsignal werden daher in bestimmten Abständen T Analyserahmen entnommen und davon die Energie $x^{\text{pow}}[k]$ berechnet:

$$x^{\text{pow}}[k] = \sum_{l=0}^{L-1} x^2(k \cdot T + l) \tag{3.10}$$

- Nutzsignal und Störung liegen meist nicht getrennt vor, sondern nur gemeinsam in Form des gestörten Signals. Ist zumindest die Rauschenergie bekannt, kann abweichend von obiger Definition statt der Nutzsignalenergie die Energie von $x(t)$ verwendet werden. Sind Signal und Störung nicht miteinander korreliert ($\sigma_{s,n} = 0$) lässt sich Gleichung (3.9) aber auch folgendermaßen umschreiben:

$$SNR = 10 \cdot \log_{10} \frac{\sigma^2_x - \sigma^2_n}{\sigma^2_n} \text{dB} = 10 \cdot \log_{10} \left(\frac{\sigma^2_x}{\sigma^2_n} - 1 \right) \text{dB} \tag{3.11}$$

[10] Sollten die Zeitsignale einen Gleichanteil (Offset) besitzen, kann man diesen leicht durch Subtrahieren des Mittelwertes der Aufnahme oder durch einen Offset-Filter ausgleichen.

Bei mittelwertfreien Signalen gilt mit $\sigma_{s,n} = 0$ auch $E(s(t) \cdot n(t)) = 0$. Ist die Summe dieses Produkts auch für die gewählten Zeitfenster in Gleichung (3.10) vernachlässigbar, gilt näherungsweise:

$$x^{\text{pow}}[k] \approx s^{\text{pow}}[k] + n^{\text{pow}}[k] \tag{3.12}$$

Damit hat man eine Beziehung gefunden, mit der sich aus Kenntnis der Energie eines Signalteils und der Gesamtenergie der jeweils andere Energiebeitrag bestimmen lässt. Bevor basierend darauf verallgemeinerte SNR-Definitionen (für die Sprachverarbeitung) vorgestellt werden können, bedarf es noch einiger grundsätzlicher Überlegungen.

- Damit der SNR-Wert nicht von der Dauer der Sprachpausen abhängt, sollte er nur für die Teile mit Sprachaktivität berechnet werden. Dazu muss diese zunächst detektiert werden, was besonders bei stark gestörten Aufnahmen schwierig ist.
- Besteht die Störung aus stationärem Rauschen (mit zeitlich gleichbleibenden statistischen Eigenschaften, siehe z.B. [Kroschel 1986, S. 26]) kann die mittlere Störsignalenergie in den Sprachpausen geschätzt werden. Statt $n^{\text{pow}}[k]$ kann dann einfach

  $$\hat{n}^{\text{pow}} = \frac{\sum_k p[k] \cdot x^{\text{pow}}[k]}{\sum_k p[k]} \tag{3.13}$$

  verwendet werden. $p[k]$ bezeichnet einen Wert zwischen 0 und 1 (evt. auch 0 oder 1), der angibt, mit welcher Sicherheit es sich bei dem Analyserahmen um Pause handelt.

  Mit der Näherung (3.12) für unkorrelierte Signale kann die reine Nutzsignalenergie $s^{\text{pow}}[k]$ geschätzt werden, indem man die mittlere Störenergie $\hat{n}^{\text{pow}}$ von der Signalenergie $x^{\text{pow}}[k]$ abzieht. Simuliert man die Störungen, dadurch dass man z.B. bestimmte Geräusche additiv dem Sprachsignal zumischt, stehen die jeweiligen Anteile zur Verfügung und brauchen nicht geschätzt zu werden.
- Da ein einzelner SNR-Wert nur ein grobes Maß für die Qualität einer Sprachaufnahme darstellt, werden zum Teil auch lokale Schätzwerte verwendet (Abschnitt 8.4.2, S. 140). Diese beziehen sich dann meist auf einzelne Zeitfenster oder spektrale Bänder des Signals (z.B. [Cappé 1994]). Da Störungen in verschiedenen Frequenzbereichen sich unterschiedlich stark bemerkbar machen, findet man auch SNR-Definitionen, die solche spektralen Schätzungen frequenzabhängig gewichtet zusammenfassen (siehe z.B. [Jayant und Noll 1984, S. 652ff], [Acero 1990, S. 23]).

Segmental-SNR

Berechnet man das Signal-zu-Rausch-Verhältnis nach Gleichung (3.9) wird der Wert durch die energiereichen Sprachsegmente dominiert. Beim „Segmental-SNR" [Noll 1974] wird zunächst ein lokales Energieverhältnis berechnet und dann über den Logarithmus gemittelt:

$$SNR_{Seg} = \frac{1}{K} \sum_{k=0}^{K-1} 10 \cdot \log_{10} \frac{s^{\text{pow}}[k]}{n^{\text{pow}}[k]} \text{ dB} \tag{3.14}$$

Diese Definition setzt die Kenntnis des Nutzsignals und des Störsignals voraus und eignet sich daher im Allgemeinen nur für Simulationen. Außerdem darf nur über Zeitfenster k gemittelt werden, für die die Signalenergie $s^{\text{pow}}[k]$ nicht verschwindet. Unter der Annahme einer stationären, nicht mit der Sprache korrelierten Störung, kann man $n^{\text{pow}}[k]$ näherungsweise durch $\hat{n}$ und $s^{\text{pow}}[k]$ durch $(x^{\text{pow}}[k]-\hat{n})$ ersetzen. Dabei können allerdings Werte ≤ 0 entstehen, die dann nicht logarithmiert werden können.

Mean-SNR

Der Definition des „Mean-SNR" (nach [Smolders *et al.* 1994]) liegt eine Sprach-Pause-Bewertung $w[k]=1-p[k]$ zugrunde. Diese kann z.B. in binärer Form durch eine Sprach-Pause-Detektion ($w[k]=1$ für Sprache und $w[k]=0$ für Pause) gegeben sein oder als Schätzung der Wahrscheinlichkeit $P(Sprache \mid \mathbf{x}[k])$ vorliegen. Die Definition lautet:

$$SNR_{Mean} = 10 \cdot \left(\frac{\sum_k w[k] \cdot \log_{10}(x^{\text{pow}}[k])}{\sum_k w[k]} - \frac{\sum_k p[k] \cdot \log_{10}(x^{\text{pow}}[k])}{\sum_k p[k]} \right) \text{dB} \qquad (3.15)$$

Es werden also gewichtete Mittelwerte der logarithmierten Energien für die Sprach- und Pausenanteile berechnet. Da die Sprachanteile auch die Störung enthalten, ergibt sich eine Sättigung bei 0 dB

Basierend auf Gleichung (3.11) lässt sich eine Näherung für ein SNR mit dem Nutzsignal finden, den sogenannten modifizierten Mean-SNR:

$$SNR_{Mod} = 10 \cdot \log\left(10^{\frac{SNR_{Mean}}{10}} - 1 \right) \text{dB} \qquad (3.16)$$

Laut [Korthauer 1999] ergibt sich ein linearer Zusammenhang (Korrelationskoeffizient = 0,91) zwischen dem modifizierten Mean-SNR und dem Segmental-SNR mit der Näherung:

$$SNR_{Mod} \approx 0{,}76 \cdot SNR_{Seg} + 3{,}48\text{dB} \quad \text{oder} \quad SNR_{Seg} \approx 1{,}32 \cdot SNR_{Mod} - 4{,}58\text{dB}$$

Bei [Smolders *et al.* 1994] ergibt sich ein ähnlicher Zusammenhang auch für den Mean-SNR:

$$SNR_{Seg} \approx 1{,}3 \cdot SNR_{Mean} - 7{,}6\text{dB}$$

Gemittelte Spektren

Um besser beurteilen zu können, wie sich das Rauschen auf einzelne Frequenzkomponenten auswirkt, schlägt [Acero 1990] vor, die gemittelten logarithmierten Sprach- und Pausenspektren darzustellen (AS von engl.: *average spectra*). Hierbei sind auch die spektralen Verzerrungen des Sprachspektrums durch Kanaleinflüsse und die Verteilung der Störung über die Frequenz zu erkennen. Die Bestimmung erfolgt ähnlich wie beim SNR_{Mean}, nur diesmal für einzelne Spektralkoeffizienten $x_i^{\text{s}}[k]$ des Leistungsspektrums:

$$AS_{Sprache}(i) = 10 \cdot \frac{\sum_k w[k] \cdot \log_{10}(x_i^*[k])}{\sum_k w[k]} \tag{3.17}$$

$$AS_{Pause}(i) = 10 \cdot \frac{\sum_k p[k] \cdot \log_{10}(x_i^*[k])}{\sum_k p[k]} \tag{3.18}$$

Die Differenz dieser beiden Funktionen lässt sich nun wiederum über alle n Koeffizienten mitteln um einen einzigen Wert zu erhalten:

$$AS_{total} = \frac{1}{n} \cdot \sum_{i=1}^{n} \left(AS_{Sprache}(i) - AS_{Pause}(i) \right) \tag{3.19}$$

Dieser Wert entspricht nicht dem SNR_{Mean}, da nun die einzelnen Koeffizienten schon vor der Summation über i logarithmiert wurden (es gilt $x^{pow}[k] = \sum_i x_i^*[k]$).

Praktische Umsetzung

Die SNR-Werte können zur reinen Beurteilung der vorhandenen oder der durch ein Verfahren verbesserten Signalqualität dienen. Mit ihnen können aber auch die Parameter eines Verfahrens abhängig vom Störpegel angepasst werden. Als Spezialfall der Anpassung können sie direkt als Merkmal zur Spracherkennung verwendet werden. Durch die unterschiedlichen Berechnungsmethoden und Abhängigkeiten (beispielsweise vom Sprach-Pause-Detektor) sagen die absoluten Werte nicht allzu viel aus. Es sollten daher immer nur relative Vergleiche mit gleicher Messmethode angestellt werden.

Bei Simulationen mit künstlich verrauschten Sprachdaten ergeben sich die Vorteile, dass die Sprache $s(t)$ und die Störung $n(t)$ bekannt sind und die Sprach-Pause-Detektion auf dem ungestörten Signal durchgeführt werden kann. Da in dieser Arbeit „echte" Daten verwendet werden, eignen sich der modifizierte Mean-SNR (S. 57) und die gemittelten Spektren (Abschnitt 8.3.1, S. 132).

3.3.2 Mittlerer Quadratischer Fehler

Kurzbeschreibung: Der mittlere quadratische Fehler beurteilt die durchschnittliche Abweichung einer Näherung oder eines Schätzwertes vom wahren Wert. Dazu wird eine Stichprobe verwendet, für die die wahren Werte bekannt sein müssen.

Definition

Der **mittlere quadratische Fehler** ***MSE*** (engl. *mean square error*) ist ein gebräuchliches Maß, um lokale Abweichungen eines Schätzwertes $\hat{x}[k]$ von einer Referenz $x[k]$ in eine globale Bewertung zu überführen. Der MSE ist der Mittelwert der quadratischen Fehler aller Beispiele einer Stichprobe:

$$MSE = \frac{1}{K} \sum_{k=1}^{K} (\hat{x}[k] - x[k])^2 \triangleq E\left\{ (\hat{x} - x)^2 \right\} \tag{3.20}$$

Bei Merkmalsvektoren $\mathbf{x}[k]$ wird die Differenz durch den Betrag des Differenzvektors ersetzt. In einer etwas abgewandelten Schreibweise kann man den MSE auch über die Spur der Fehler-Korrelationsmatrix $\mathbf{S}_{e,e}$ ausdrücken (ähnlich wie bei der Linearen Diskriminanz im nächsten Abschnitt):

$$MSE = \frac{1}{K}\sum_{k=1}^{K}\left|\hat{\mathbf{x}}[k]-\mathbf{x}[k]\right|^2 \triangleq \mathrm{tr}\left(\mathbf{S}_{e,e}\right) = \mathrm{tr}\left(\mathrm{E}\left\{\mathbf{e}\cdot\mathbf{e}^{\mathrm{T}}\right\}\right) \tag{3.21}$$

mit dem Fehlervektor $\mathbf{e}[k] = \hat{\mathbf{x}}[k]-\mathbf{x}[k]$

Die Korrelationsmatrix $\mathbf{S}_{e,e}$ ist dabei der Erwartungswert von $\mathbf{e}\cdot\mathbf{e}^{\mathrm{T}}$.

Verwendung als Optimierungskriterium

Der *MSE* wird zur Problemlösung in der Klassifikation, der Parameter- und der Signalschätzung eingesetzt. Man spricht dann meist von einem **MMSE-Verfahren** (engl. *minimum mean square error*). Abstrahiert man von den unterschiedlichen Anforderungen und verwendet eine vereinfachte Darstellung, so beruht die gesuchte Ausgabe (Schätzwert $\hat{x}$) stets auf einer Eingabe $y = f(x)$, die aus dem Referenzwert x hervorgegangen ist. Der optimale Ausgabewert im Sinne des MSE-Kriteriums ist dabei durch den Erwartungswert von x unter der Bedingung y gegeben. Obwohl es sich dabei nicht unbedingt um den wahrscheinlichsten Wert handelt (siehe **Abbildung 3.3**), ist für diesen Schätzwert der quadratische Fehler im Mittel am geringsten:

$$\mathrm{E}\{MSE \mid y\} = \int_x (\hat{x}-x)^2\, \mathrm{p}(x \mid y)dx$$

$$\frac{\delta}{\delta\hat{x}}\mathrm{E}\{MSE \mid y\} = \frac{\delta}{\delta\hat{x}}\int_x (\hat{x}-x)^2\, \mathrm{p}(x \mid y)dx = \int_x \frac{\delta}{\delta\hat{x}}(\hat{x}-x)^2\, \mathrm{p}(x \mid y)dx$$

$$= \int_x 2(\hat{x}-x)\mathrm{p}(x \mid y)dx = 2\left(\hat{x}-\mathrm{E}\{x \mid y\}\right) \overset{!}{=} 0$$

$$\Rightarrow \quad \hat{x} = \mathrm{E}\{x \mid y\} \tag{3.22}$$

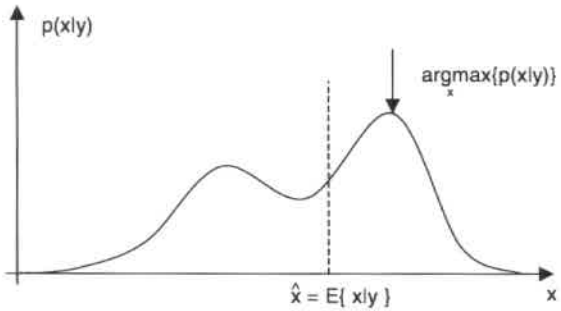


Abbildung 3.3: Schätzwert mit MMSE und nach Maximum-a-posteriori-Prinzip (MAP).

Für die Berechnung einer Klassifikatorfunktion ergeben sich mit dem MSE-Kriterium erhebliche Abweichungen gegenüber der MAP-Bestimmung (Maximum-a-posteriori), sofern es sich nicht um sehr einfache Klassenverteilungen handelt [Dülfer 1993, S. 54]. Daher ist die Verwendung eher bei der Parameterschätzung üblich. Dort stimmen die Schätzwerte nach MMSE und MAP für Verteilungen überein, bei denen Mittelwert und Maximum zusammenfallen (wie z.B. bei Gaußdichten [Kroschel 1986, S. 136]).

In der Spracherkennung wird der MSE vor allem im Zusammenhang mit der Schätzung ungestörter Merkmale $\mathbf{x}$ basierend auf gestörten Merkmalen $\mathbf{y}$ verwendet. Dabei ist die bedingte Dichte $p(\mathbf{x}|\mathbf{y})$ meist nicht bekannt. D.h. man benötigt eine *verteilungsfreie* Schätzeinrichtung. Es wird ein Merkmalsvektor $\mathbf{x}[k]$ gesucht, für den nur ein abgeleiteter Merkmalsvektor $\mathbf{y}[k] = f(\mathbf{x}[k])$ zur Verfügung steht. Der Vektor $\mathbf{y}[k]$ kann durch eine Störung entstanden sein oder ist generell durch eine unbekannte Abbildung $f(..)$ aus $\mathbf{x}[k]$ hervorgegangen. Nun soll eine Transformation gefunden werden, die diese Abbildung rückgängig macht und so basierend auf $\mathbf{y}[k]$ einen Schätzwert $\hat{\mathbf{x}}[k]$ für das unbekannte $\mathbf{x}[k]$ findet. Mittels einer Modellvorstellung für die Abbildung $f(..)$ wird eine geeignete Umkehrfunktion definiert, deren Parameter gefunden werden sollen.

Dazu können eine Stichprobe mit korrespondierenden Wertepaaren $\mathbf{x}[k]$ und $\mathbf{y}[k]$ oder zumindest daraus abgeleitete, statistische Zusammenhänge verwendet werden. Für lineare Schätzer $\hat{\mathbf{x}}[k] = \boldsymbol{\mu}_x + \mathbf{A} \cdot (\mathbf{y}[k] - \boldsymbol{\mu}_y)$ ergibt sich eine geschlossene Lösung mit dem Gauß-Markoff-Theorem zur Bestimmung von $\mathbf{A}$:

$$\mathbf{A} = \mathbf{C}_{x,y} \cdot \mathbf{C}_{y,y}^{-1} \qquad \text{mit} \quad \mathbf{C}_{x,y} = \mathrm{E}\{(\mathbf{x} - \boldsymbol{\mu}_x) \cdot (\mathbf{y} - \boldsymbol{\mu}_y)^{\mathrm{T}}\} \qquad (3.23)$$

$$\text{und} \quad \mathbf{C}_{y,y} = \mathrm{E}\{(\mathbf{y} - \boldsymbol{\mu}_y) \cdot (\mathbf{y} - \boldsymbol{\mu}_y)^{\mathrm{T}}\}$$

Eine Herleitung findet man in [Kroschel 1986, S. 164] (abweichend davon wurden in der oben verwendeten Darstellung die Mittelwerte berücksichtigt). Ähnliche Lösungen werden für die Filterung von gestörten Signalen verwendet (Wiener-Filter) oder zur Rekonstruktion unzuverlässiger Koeffizienten in der *Missing Feature Theory* (siehe Abschnitt 5.2).

Der MSE wird auch zur Bestimmung der Parameter von Näherungsfunktionen eingesetzt. Ein schon erwähntes Beispiel aus dem Bereich Spracherkennung ist die Berechnung von *Delta-Koeffizienten* in der Merkmalsextraktion (siehe Seite 22). Dort verwendet man die zum aktuellen Sprachrahmen k benachbarten Merkmalsvektoren als Stichprobe. Über das MMSE-Prinzip werden die Parameter einer Polynomapproximation zweiten Grades bestimmt. Deren Ableitung wird dann schließlich als dynamisches Merkmal verwendet.

Verwendung als Gütemaß

Handelt es sich bei $\mathbf{y}$ um eine gestörte Version von $\mathbf{x}$ (d.h. sind es Muster im selben Merkmalsraum), kann man dessen mittlere Abweichung mithilfe des quadratischen Fehlers bestimmen. Durch Vergleich des quadratischen Fehlers für einzelne Merkmale lässt sich auch feststellen, welche der Merkmale besonders gestört sind.

Wird durch ein Verfahren mit einem beliebigen Optimierungskriterium die Störung reduziert, so sollte auch der *MSE* für den neuen Merkmalsvektor $\hat{\mathbf{x}}[k]$ geringer sein. Ähnlich

wie für die Wortfehlerratenreduktion (siehe Abschnitt 3.1) definieren wir daher die Reduktion des MSE als:

$$MSE_{Red} = 1 - \frac{\sum_k \left\|\hat{\mathbf{x}}[k] - \mathbf{x}[k]\right\|^2}{\sum_k \left\|\mathbf{y}[k] - \mathbf{x}[k]\right\|^2} \tag{3.24}$$

Diese Reduktion kann für beliebige Merkmalsräume, die auf den verbesserten Merkmalen $\hat{\mathbf{x}}[k]$ aufbauen, betrachtet werden. Da zum Beispiel Geräuschreduktionsfilter meist im Spektralbereich arbeiten, kann man so auch deren Auswirkung im Cepstralbereich ermitteln (Cepstrale Distanz, Abschnitt 8.4.3, S. 145). Dort ist aber der quadratische Fehler nicht immer sinnvoll, denn die einzelnen Merkmale haben stark unterschiedliche Varianzen und sind unterschiedlich wichtig. Um dies auszugleichen wird eine normierte Reduktion eingeführt. Dazu werden die Schwankungen des ungestörten Merkmalsvektors über die Stichprobe (z.B. eine Äußerung) betrachtet. Es wird angenommen, dass in der Varianz der Merkmale die gewünschte Information steckt. Die quadratischen Fehler des gestörten Wertes und des Schätzwertes werden vor der Zusammenfassung durch diese Varianz geteilt. So ergibt sich ein normierter MSE (NMSE) und damit die entsprechende Reduktion zu:

$$NMSE_{Red} = 1 - \frac{\sum_k \sum_{i=1}^{N} \frac{\left(\hat{x}_i[k] - x_i[k]\right)^2}{\sigma^2_{x_i}}}{\sum_k \sum_{i=1}^{N} \frac{\left(y_i[k] - x_i[k]\right)^2}{\sigma^2_{x_i}}} \tag{3.25}$$

3.3.3 Lineare Diskriminanz

***Kurzbeschreibung**: Die lineare Diskriminanz beurteilt die Unterscheidbarkeit von Klassen anhand einer Stichprobe für die diese Klassenzuteilung gegeben ist. Je größer ihr Wert, umso leichter kann ein linearer Klassifikator die Klassen unterscheiden.*

Die **Lineare Diskriminanz** (LD) ist kein fest definierter Begriff. Im Rahmen dieser Arbeit wird eines von mehreren alternativen Optimierungskriterien aus der Diskriminanzanalyse [Fukunaga 1972, S. 261] als Lineare Diskriminanz bezeichnet. Dort werden Kriterien verwendet, mit denen man die Unterscheidbarkeit von vorgegebenen Klassen beurteilen möchte. Die Verteilung der Muster innerhalb der Klassen wird als relativ gutmütig angenommen, was heißen soll: die Wahrscheinlichkeitsdichten lassen sich näherungsweise mit Gaußdichten beschreiben. Die Unterscheidbarkeit ist dann umso größer

- je näher die Muster einer Klasse an ihrem Klassen-Erwartungswert liegen und
- je weiter die Erwartungswerte der Klassen oder die Muster insgesamt auseinander liegen.

Zur Beschreibung dieser Streuungen werden zunächst folgende Matrizen definiert (in Klammern sind die jeweils üblichen englischen Bezeichnungen angegeben, aus denen auch die Benennung der Matrizen hervorgeht):

- Klassen-Kovarianzmatrix mit den Klassen $s = 1 \ldots S$:
$$\mathbf{W}_s = \mathrm{E}\left\{(\mathbf{x}-\boldsymbol{\mu}_s)\cdot(\mathbf{x}-\boldsymbol{\mu}_s)^{\mathrm{T}} \mid \mathbf{x} \in s\right\} \qquad \text{mit } \boldsymbol{\mu}_s = \mathrm{E}\left\{\mathbf{x} \mid \mathbf{x} \in s\right\} \tag{3.26}$$

- Mittlere Klassen-Kovarianzmatrix (engl.: *within-class scatter*):
$$\mathbf{W} = \sum_{s=1}^{S} \mathrm{P}(s)\cdot\mathbf{W}_s \tag{3.27}$$

- Kovarianzmatrix der Mittelwerte (engl.: *between-class scatter*):
$$\mathbf{B} = \sum_{s=1}^{S} \mathrm{P}(s)\cdot\left\{(\boldsymbol{\mu}_s-\boldsymbol{\mu})\cdot(\boldsymbol{\mu}_s-\boldsymbol{\mu})^{\mathrm{T}}\right\} \qquad \text{mit } \boldsymbol{\mu} = \mathrm{E}\{\mathbf{x}\} \tag{3.28}$$

- Kovarianzmatrix (engl.: *total scatter*):
$$\mathbf{T} = \mathrm{E}\left\{(\mathbf{x}-\boldsymbol{\mu})\cdot(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\right\} = \mathbf{W}+\mathbf{B} \tag{3.29}$$

Abbildung 3.4 zeigt beispielhaft zwei Arten dieser Kovarianzmatrizen ($\mathbf{T}$ und $\mathbf{W}_s$), dargestellt als Ellipsen. Die Punkte einer Ellipse beschreiben eine Höhenlinie der entsprechenden Gaußdichte, d.h. sie liefern konstante Werte bei der Auswertung der Dichte. Die Richtung der Hauptachsen entspricht den Eigenwerten der Kovarianzmatrix und die Länge den zugehörigen Eigenwerten (in der Abbildung als $\mathbf{e}_1$ und $\mathbf{e}_2$ für einen Fall angedeutet).

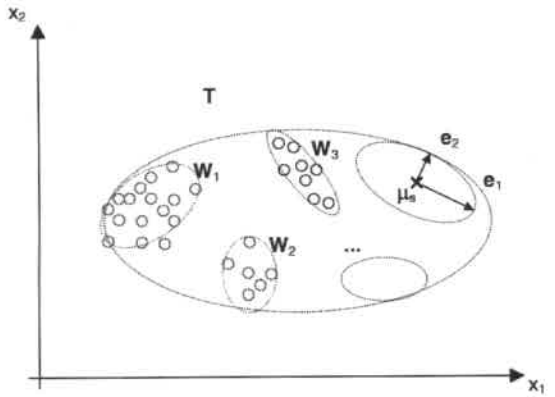


Abbildung 3.4: Muster verschiedener Klassen und Kovarianzmatrizen zur Beschreibung der Streuung aller Muster ($\mathbf{T}$) und einzelner Klassen ($\mathbf{W}_s$).

Um ein Kriterium für die Unterscheidbarkeit der Klassen zu formulieren, muss zunächst aus den Matrizen eine Größe abgeleitet werden, die die jeweilige Streuung der betrachteten Muster mit einem Wert beschreibt. Dazu eignet sich beispielsweise die Spur der Kovarianzmatrix. Sie ergibt sich aus der Summe der Diagonalelemente, was wiederum nichts anderes ist als der mittlere quadratische „Fehler" (euklidischer Abstand) zum Mittelwert oder auch die Summe der Eigenwerte.

- Spur, am Beispiel der Kovarianzmatrix **T** mit Dimension *n*:

$$\mathrm{tr}(\mathbf{T}) = \sum_{i=1}^{n} \sigma_i^{\,2} = \sum_{i=1}^{n} \mathrm{E}\left\{(x_i - \mu_i)^2\right\} = \mathrm{E}\left\{\sum_{i=1}^{n} (x_i - \mu_i)^2\right\} = \mathrm{E}\left\{\left|\mathbf{x} - \boldsymbol{\mu}\right|^2\right\} \tag{3.30}$$

Ebenfalls geeignet und auch im Folgenden verwendet ist die Determinante, die als Produkt der Eigenwerte in der Linearen Diskriminanzanalyse zur selben Lösung des dort gestellten Problems führt wie über die Spur definierte Kriterien. In [Fukunaga 1972] werden insgesamt vier Kriterien mit Spur oder Determinante vorgestellt, die zusätzlich eine alternative Verwendung und verschiedene Kombinationen der Kovarianzmatrizen erlauben.

Für die vorliegende Arbeit wird die **Lineare Diskriminanz** wie folgt festgelegt:

$$LD = \left|\mathbf{W}^{-1} \cdot \mathbf{T}\right| = \frac{|\mathbf{T}|}{|\mathbf{W}|} \tag{3.31}$$

Sie setzt also die Streuung sämtlicher Muster ins Verhältnis zur mittleren Streuung der Klassen. In der Spracherkennung können diese Muster je nach Merkmalsextraktion unterschiedlich aussehen. Ist der Wert für *LD* groß, handelt es sich im Allgemeinen um geeignete Merkmale, die eine leichtere Unterscheidung der Klassen zulassen.

3.3.4 Likelihood

***Kurzbeschreibung**: Die Likelihood (wie sie hier verwendet wird) beurteilt, wie wahrscheinlich eine Stichprobe von Merkmalsvektoren mit gegebener Klassenzuordnung für ein gegebenes Modell ist. Ist ihr Wert groß, so handelt es sich um typische Beispiele für diese Klassen.*

Likelihood bei der Klassifikation

Die Likelihood dient als Entscheidungskriterium für einen Klassifikator, wenn keine A-priori-Informationen gegeben sind. Angenommen man möchte zufällig herausgegriffene Merkmalsvektoren $\mathbf{x}[k]$ den phonetischen Klassen z_s eines zuvor trainierten akustischen Modells zuordnen (Zuordnung $z[k]$). Dazu würde man möglichst einen Bayes-Klassifikator verwenden, der die A-posteriori-Wahrscheinlichkeit $\mathrm{p}(z_s|\mathbf{x}[k])$ maximiert (Maximum-a-posteriori, vergleiche auch Gleichung (2.1)):

$$\text{MAP:} \qquad z[k] = \arg\max_s \left(\mathrm{P}(z_s|\mathbf{x}[k])\right) = \arg\max_s \left(\mathrm{p}(\mathbf{x}[k]|z_s) \cdot \mathrm{P}(z_s)\right) \tag{3.32}$$

Dabei liefert das akustische Modell, das zum Beispiel aus Gauß'schen Mischverteilungen für jede Klasse bestehen könnte, eine Schätzung der bedingten Dichte $\mathrm{p}(\mathbf{x}[k]|z_s)$. Die A-priori-Wahrscheinlichkeiten $\mathrm{P}(z_s)$ könnten in diesem Fall aus den relativen Häufigkeiten der phonetischen Klassen bestimmt werden. Liegt letztere Information nicht vor, muss man sich auf die Auswertung der sogenannten Likelihood, der bedingten Dichte $\mathrm{p}(\mathbf{x}[k]|z_s)$, verlassen (Maximum-Likelihood):

$$\text{ML:} \qquad z[k] = \arg\max_s \left(\mathrm{p}(\mathbf{x}[k]|z_s)\right) \tag{3.33}$$

Likelihood bei der Parameterschätzung

Die Likelihood wird in dieser Arbeit hauptsächlich in einem speziellen Zusammenhang zur Parameterschätzung verwendet. Dabei wird angenommen, dass die gestörten Merkmalsvektoren $\tilde{\mathbf{x}}[k] = f(\mathbf{x}[k], \tilde{\alpha})$ und eine Klassenzuordnung $z[k]$ gegeben sind. Die Klassen entsprechen dabei entweder den Zuständen eines Satz-HMMs oder den Gaußdichten einer Mischverteilung, die als akustisches Modell λ die ungestörten Muster $\mathbf{x}$ beschreiben. Mithilfe der Likelihood sollen die Störparameter $\tilde{\alpha}$ geschätzt werden, damit Muster und Modell wieder besser zueinander passen. Dazu verändert man entweder

- die Modellparameter $\tilde{\lambda} = \Phi(\lambda, \alpha)$, damit sie an die andersartigen Merkmale angepasst werden oder
- die Merkmale $\hat{\mathbf{x}} = g(\tilde{\mathbf{x}}, \alpha)$, so dass die Merkmalsvektoren besser zum Modell passen.

Die Umkehrfunktion $g(\ldots, \alpha)$ (oder auch Schätzfunktion) wird meist aufgrund einer Modellvorstellung über die Störfunktion $f(\ldots, \alpha)$ festgelegt. Entsprechendes gilt für die Modellanpassungsfunktion $\Phi(\ldots, \alpha)$.

Die *Log*-Likelihood (hier auch kurz Likelihood genannt) ist dabei definiert als der Logarithmus der Wahrscheinlichkeit einer Folge von Merkmalsvektoren $\mathbf{X} = \mathbf{x}[1] \ldots \mathbf{x}[K]$ für ein gegebenes Modell λ. Setzt man die Ereignisse als unabhängig voraus, ergibt sich:

$$\textit{Log}\text{-Likelihood:} \quad \log\left(p(\mathbf{X}|\lambda)\right) = \log\left(\prod_k p(\mathbf{x}[k]|\lambda)\right) = \sum_k \log\left(p(\mathbf{x}[k]|\lambda)\right) \tag{3.34}$$

Mit den Zuordnungen $z[k]$ lässt sich nun eine vereinfachte Auswertung vornehmen, mit der man geeignete Parameter $\hat{\alpha}$ finden kann. Die optimale Einstellung erhält man über

$$\hat{\alpha} = \arg\max_{\alpha}\left(\sum_k \log\left(p(\tilde{\mathbf{x}}[k]\,|\,z[k], \Phi(\lambda, \alpha))\right)\right) \quad \text{für Parameter des Modells} \tag{3.35}$$

$$\hat{\alpha} = \arg\max_{\alpha}\left(\sum_k \log\left(\tilde{p}(g(\tilde{\mathbf{x}}, \alpha)|\, z[k], \lambda)\right)\right)$$

für Parameter der Merkmalsextraktion. (3.36)

Im zweiten Fall geht man zunächst ebenfalls von einer Likelihood-Bewertung für die gestörten Merkmalsvektoren $\tilde{\mathbf{x}}[k]$ mittels eines entsprechenden Modells $\tilde{\lambda}$ aus. Diese Bewertung kann unter Umständen über eine Modelltransformation (siehe Seite 91) aus dem gegebenen Modell λ und einem Schätzwert $\hat{\mathbf{x}} = g(\tilde{\mathbf{x}}, \alpha)$ für die ungestörten Muster $\mathbf{x}$ berechnet werden (hier angedeutet durch $\tilde{p}(\ldots)$). In einigen Fällen (wie bei der Sprechernormierung mit ML-VTLN, siehe Abschnitt 7.4) wird auch direkt der Schätzwert $\hat{\mathbf{x}}$ mit dem Modell λ ausgewertet:

$$\hat{\alpha} = \arg\max_{\alpha}\left(\sum_k \log\left(p(g(\tilde{\mathbf{x}}, \alpha)|\, z[k], \lambda)\right)\right) \tag{3.37}$$

Anzumerken ist noch, dass die Klassenzuordnung $z[k]$ in den später benutzten Anwendungsfällen teilweise von den ursprünglichen Merkmalen oder dem Modell abhängt.

Durch eine ML-Parameterwahl $\hat{\alpha}$, die von den ursprünglich verwendeten Parametern abweicht, ergibt sich dann meist eine ebenfalls abweichende Zuordnung. Durch Iteration kann man in diesen Fällen Parameter und Zuordnung abwechselnd optimieren bis ein bestimmtes Abbruchkriterium erreicht wird.

Genau dieses Prinzip wird auch beim Training der Parameter der akustischen Modelle verfolgt. Bei gegebener Zuordnung $\gamma_{s,m}[k]$ der Merkmalsvektoren $\mathbf{x}[k]$ zu den Zuständen z_s ($s = 1..S$) und darüber hinaus zu den einzelnen Gaußdichten m werden die Parameter über das ML-Kriterium neu bestimmt [Duda und Hart 1973, S. 49].

Mittelwert:
$$\hat{\mu}_{s,m} = \frac{1}{\sum_k \gamma_{s,m}[k]} \sum_k \gamma_{s,m}[k] \cdot \mathbf{x}[k] \tag{3.38}$$

Kovarianzmatrix:
$$\hat{\mathbf{C}}_{s,m} = \frac{1}{\sum_k \gamma_{s,m}[k]} \sum_k \gamma_{s,m}[k] \cdot \left(\mathbf{x}[k] - \hat{\mu}_{s,m}\right) \cdot \left(\mathbf{x}[k] - \hat{\mu}_{s,m}\right)^{\mathrm{T}} \tag{3.39}$$

Mixturgewicht:
$$\hat{\mathrm{P}}(m \mid z_s) = \frac{1}{\sum_{m \in s} \sum_k \gamma_{s,m}[k]} \sum_k \gamma_{s,m}[k] \tag{3.40}$$

Anfangs- und Übergangswahrscheinlichkeiten ergeben sich dabei durch Bestimmen der relativen Häufigkeiten. Für die Schätzwerte der Iteration i wird die Zuordnung $\gamma_{s,m}[k]$ des Vorgängermodells λ_{i-1} über den Forward-Backward- oder den Viterbi-Algorithmus ermittelt (vergleiche Abschnitt 2.4).

4 Experimenteller Rahmen

Kapitelübersicht: Dieses Kapitel beschreibt die praktischen Grundlagen der vorliegenden Arbeit. Dazu gehören erstens Sprachdaten, die zum Trainieren und Testen eines Spracherkenners nötig sind (Abschnitt 4.1), zweitens ein Ausgangssystem (Abschnitt 4.2: Basissystem), das soweit auf die Anforderungen der hier betrachteten Anwendung optimiert ist, dass damit Untersuchungen zur Robustheit in stark gestörten und wechselnden Umgebungen durchgeführt werden können.

4.1 Sprachdaten

Eine der wichtigsten Voraussetzung zur Erstellung eines automatischen Spracherkenners sind geeignete Trainingsdaten in möglichst großer Menge. Diese Daten müssen verschriftet (transliteriert) sein, d.h. der gesprochene Wortlaut als Text zur Verfügung stehen. Am einfachsten erhält man solche Daten, indem man Sprecher (Sprachspender) akquiriert, die ihnen vorgelegte Texte oder auch gewünschte Kommandowörter vorlesen. Diese Daten können dann zum Trainieren eines Diktiererkenners bzw eines Kommandoworterkenners benutzt werden. In einem gewissen Umfang ist man in Hinblick auf den späteren Einsatz unabhängig von den verwendeten Texten oder Wörtern, denn trainiert werden meist phonetische Einheiten, die wiederum neu zu anderen Wörtern zusammengesetzt werden können. Dazu ist für jedes Wort der Trainingsmenge eine phonetische Umschrift nötig oder anders ausgedrückt: ein Eintrag im Aussprachewörterbuch. Gerade beim Beispiel der Kommandowörter leuchtet es jedoch ein, dass man schon beim Training möglichst Wörter aus der späteren Anwendung verwenden sollte, um auch Koartikulationseffekte zu erfassen. Ist diese Voraussetzung erfüllt, spricht man von *domänen-spezifischen* Trainingsdaten. Bei kleinem Erkennerwortschatz (Vokabular) ist dann eventuell sogar eine Ganzwortmodellierung sinnvoll, bei der jedes Wort ein eigenes Modell aus einer Reihe von Zuständen zugeordnet bekommt. Um diese Wortmodelle zu trainieren, braucht man keine phonetische Umschrift, aber genügend Trainingsbeispiele für jedes Wort.

Sehr viel schwieriger ist die Beschaffung *spontansprachlicher* Anfragen. Diese werden während der Benutzung eines bereits existierenden Systems gesammelt oder mithilfe eines entsprechend gestalteten Szenarios aus den Spendern „herausgelockt". Anschließend müssen die so gewonnenen Audiodaten mehrfach angehört und verschriftet werden. Diese aufwändige Beschaffungsmaßnahme hat hauptsächlich zwei Gründe:

1. Die Texte, die sich aus der Verschriftung ergeben, werden zur Erstellung des (linguistischen) Sprachmodells benötigt. Spontane Sprache unterscheidet sich in Wortwahl und Grammatik stark von gelesener Sprache, es liegen dafür aber kaum Beispiele in schriftlicher Form vor.

2. Die akustische Realisierung spontaner Sprache unterscheidet sich von gelesener Sprache, da verstärkt Füllwörter, Verschleifungen (besonders bei oft benutzten Phrasen), unterschiedliche Sprechgeschwindigkeiten und dergleichen vorkommen. Mithilfe der Verschriftungen und den Audioaufnahmen wird das akustische Modell trainiert.

Der zweite Punkt beinhaltet auch, dass sich die Realisierung der Wörter seltener an die Standardaussprache des Wörterbuches hält. Um dies zu kompensieren, werden weiter unten noch Maßnahmen beschrieben.

4.1.1 Spontansprachliche Trainingsdaten aus der Laborumgebung (GSST)

Große Mengen an verschrifteten Sprachdaten stehen heutzutage vor allem für gelesene Sprache im Labor sowie für spezielle Anwendungen mit Einzelwörtern oder Ziffernketten zur Verfügung. Dies trifft natürlich nur für solche Sprachen zu, für die auch intensiv an der automatischen Spracherkennung gearbeitet wird. Im Rahmen des Verbundprojektes VerbMobil, das vom BMBF (Bundesministerium für Bildung und Forschung) gefördert wird, wurde eine große Menge spontansprachlicher, deutscher Dialoge aus der Domäne Termin- und Reiseplanung erstellt. Für diese Domäne wurden auch in anderen Sprachen Daten gesammelt, weshalb hier die deutschsprachige Variante als **GSST** (German Spontaneous Scheduling Task) bezeichnet wird.

GSST-Daten dienten in dieser Arbeit als Trainingsmaterial eines spontansprachlichen Erkenners für mobile Informationssysteme. Die Domäne (Anfragen an ein Informationssystem) weicht damit etwas ab, die Sprechweise ist aber ebenfalls spontan. Alle verwendeten Sprachdaten wurden in ruhiger Laborumgebung mit einem qualitativ hochwertigen Nahbesprechungsmikrofon aufgenommen. Neben regulären Wörtern wurden auch Geräusche und Pausen verschriftet, so dass für diese eigene Modelle trainiert werden können. So werden Einfügefehler reduziert und eventuell wichtige Informationen für die anschließende Auswertung der spontansprachlichen Anfrage geliefert.

Die verwendete Trainingsmenge besteht aus etwa 14.400 Äußerungen von 1.750 Sprechern und hat eine Gesamtdauer von 30 Stunden. Das zum Trainieren benutzte Aussprachewörterbuch zählt über 10.000 verschiedene Einträge.

4.1.2 Test- und Trainingsdaten aus der Autoumgebung

Für Anwendungen im Auto konzentriert man sich derzeit, wie schon in der Einführung beschrieben, auf Erkenner für Kommandowörter für den „Prompted Approach" (siehe Seite 9). In diesem Zusammenhang sollte man den Begriff Kommandowort allerdings etwas ausdehnen, da neben reinen Kommandowörtern auch Kommandophrasen, Namen, Buchstaben und Ziffern verwendet werden. Besonders bei großen Namenlisten, wie sie bei Informationssystemen immer wieder vorkommen können (Städte, Straßen, Personennamen), spielen Buchstabierungen eine wichtige Rolle [Van Compernolle 1997]. Sie reduzieren das Vokabular bei der Selektion eines Eintrages aus einer großen Liste auf die Buchstaben des Alphabets. Durch einen Buchstabiererkenner oder durch Kombination mit ihm lassen sich

Einträge aus großen Listen besser erkennen als nur mit einem Erkenner für fließend ausgesprochene Namen [Hild 1997]. Auch bei spontansprachlichen Anfragen tauchen hin und wieder einzelne oder Sequenzen von Buchstaben auf. Ein Beispiel dafür ist: „Ich möchte nach Homburg, mit ‚O'!" oder „..., das schreibt sich H, O, M, B, U, R, G". Ziffern hingegen sind hauptsächlich zur Eingabe von Telefonnummern wichtig, können bei einem Navigationssystem aber auch zur Angabe der Postleitzahl dienen.

[Langmann *et al.* 1998] gibt eine Übersicht bisher veröffentlichter Datensammlungen verschiedener Sprachen in der Autoumgebung und stellt gleichzeitig die umfangreiche Sammlung des BMBF-Projektes MoTiV (Mobilität und Transport im intermodalen Verkehr) vor. Hieran beteiligten sich neben Universitäten und Forschungsabteilungen, die sich allgemein mit Spracherkennung beschäftigen, auch namhafte Firmen aus der Automobilindustrie. In allen dort beschriebenen Sammlungen liegt die Konzentration auf Kommandowörter, Namen, Ziffern und Buchstaben. Zwar wurden in MoTiV auch einige kontinuierliche Sätze gesammelt, diese dienten aber hauptsächlich zur phonetischen Abdeckung (für zusätzliche Kommandowörter oder Namen). Erst in jüngster Zeit startete mit dem EU-Projekt SpeechDat-Car (LE4-8334) eine groß angelegte Sammlung von Sprachendaten im Auto für mehrere europäische Sprachen, was das große Interesse an Daten dieser Art bestätigt. Die Zusammensetzung der Äußerungstypen ist wiederum ähnlich wie bei MoTiV. In [Draxler *et al.* 1999] werden erste Erfahrungen mit der Sammlung für Deutsch berichtet, die als Testbett für die Aufnahmesoftware des Projektes diente.

Da das angestrebte Ziel der vorliegenden Arbeit ein kontinuierlicher Erkenner für realistische spontansprachliche Anfragen in der Autoumgebung war, wurde aus Ermangelung entsprechender Sprachdaten eine eigene Sammlung durchgeführt. Die folgenden Grundlagen sollten damit geschaffen werden:

1. Ein **Testset** mit spontansprachlichen Anfragen an ein Autonavigationssystem. Dieses soll Aufschlüsse über die Leistungsfähigkeit eines Spracherkenners unter unterschiedlichen Umgebungsbedingungen (wie Geschwindigkeit usw.) geben.

2. Ein **Trainingsset** mit kontinuierlicher Sprache aus der Autoumgebung mit dem ein Vergleichssystem trainiert werden kann. Dieses soll zeigen, welche Erkennungsergebnisse erzielt werden können, wenn dafür Trainingsdaten aus dieser speziellen Umgebung zur Verfügung stehen, auch wenn dies mit dem hohen Aufwand einer Sammlung verbunden ist.

3. Reine **Geräuschaufnahmen** aus der Autoumgebung, um durch künstliches Zumischen deren Einfluss auf typische Merkmale der Spracherkennung untersuchen zu können. Außerdem können auf diese Art auch aus Laboraufnahmen verrauschte Daten simuliert werden, die den realistischen Daten zumindest in Teilaspekten ähneln.

Im Folgenden werden einige Einzelheiten der Sammlung dargestellt, die für die später vorgestellten Experimente relevant sind. Weitere Details sind in [Westphal 1998] festgehalten.

Aufnahmeplattform

Eine weit verbreitete Möglichkeit Sprachdaten in digitaler Form zu sammeln, stellen Dat-Rekorder dar. Aufgrund von Erfahrungen bei einer vorherigen Sammlung wurde diesmal ein tragbarer Rechner zur Aufnahme der Äußerungen bevorzugt. Dies bietet mit einem entsprechenden Sammelprogramm einige Vorteile. Neben einer automatischen Qualitätskontrolle der Aufnahmen können so die Umgebungsbedingungen für jede Äußerung festgehalten und später automatisch ausgewertet werden (s.u.). Dadurch, dass die Aufnahme vom Sprecher durch Knopfdruck selbst initiiert wurde, ergab sich auch ein anwendungsnäheres Szenario als bei einem ständig mitlaufenden Band. Die dadurch bereits segmentierten Äußerungen verringerten den Aufwand bei der Nachbearbeitung der Daten. Die Abtastrate wurde wie bei den GSST-Trainingsdaten zu 16 kHz und die Quantisierung zu 16 bit gewählt.

Mikrofone

Alle Äußerungen wurden simultan mit zwei Mikrofonen aufgenommen. Das eine, ein Kondensatorwandler (AKG C400), wurde auch bei der MoTiV-Datensammlung verwendet. Es war fest im Fahrzeug über der Windschutzscheibe installiert (für eine Untersuchung verschiedener Mikrofonpositionen siehe [Smolders *et al.* 1994]) und stellt den realistischen Fall in einer späteren Anwendung dar. Das andere, ein Druckgradientenempfänger (Sennheiser HMD 410), ist ein mit Kopfhörer kombiniertes Nahbesprechungsmikrofon, das auch für die Sammlung der GSST-Trainingsdaten benutzt wurde. Es nimmt daher weit weniger Störungen auf als das fest installierte Mikrofon und die Aufnahmesituation stimmt von den Kanaleigenschaften besser mit den GSST-Daten überein. Da man es dem Fahrer oder Beifahrer bei einer kommerziellen Anwendung nicht zumuten möchte vor einer Anfrage erst ein Nahbesprechungsmikrofon „anzulegen", dient es nur zu Referenzzwecken.

Freie Szenarien

Es wurden mehrere Szenarien zur Sammlung verwendet. Mit dem Begriff *„freie Szenarien"* ist gemeint, dass die äußeren Bedingungen nicht festgelegt wurden. Geschwindigkeit und Strecke wurden nicht beschränkt, wohl aber protokolliert, um spätere Auswertungen zu ermöglichen (siehe **Tabelle 4.1**). Es wurden hierfür über 10.000 Äußerungen mit einer Gesamtdauer von 12½ Stunden gesammelt. 86 % der Äußerungen sind kontinuierlich, der Rest beinhaltet einige isoliert gesprochene Wörter sowie fast[11] reine Buchstabiersequenzen. Ein kleinerer Anteil der kontinuierlichen Äußerungen wurde durch ein Navigationsszenario initiiert, es handelt sich also hierbei um domänenspezifische, spontansprachliche Daten. Um die aufwändige Verschriftung zu reduzieren, besteht der größte Teil der kontinuierlichen Daten jedoch aus vorgelesenen Texten.

[11] Die Buchstabierungen enthalten einige interessante Effekte. So wurden, neben typischen Abweichungen wie *„doppel L"* oder *„A Umlaut"*, die Endungen gesondert behandelt: *„...straße"* wurde entweder gar nicht gesprochen, zum Teil mit *„doppel S"* buchstabiert oder einfach fließend gesprochen.

Tabelle 4.1: Protokollierte Bedingungen der freien Szenarien

Fahrsituation	Geschwindigkeit	Straße	Lüftung	Fenster	Wetter	Kommentar
Motor aus Stadtverkehr Landstrasse Autobahn	0 0-30 30-60 60-90 90-120 >120	normal Tunnel Schotter Kopfstein	aus schwach mittel stark	geschlossen offen	trocken Niesel Regen(Scheibenwischer) nass	Blinker bergauf Unebenheit vorbeifahrendes Auto Straßenbahn Motorrad beschleunigen Tanken reines Fahrgeräusch

Abbildung 4.1 und **Abbildung 4.2** geben einen Eindruck über die Häufigkeiten innerhalb der Bedingungen *„Fahrsituation"* und *„Geschwindigkeit"* (siehe auch obige Tabelle). Über die Hälfte der Aufnahmen wurde im Stadtverkehr gesammelt, der Rest verteilt sich auf Landstraße und Autobahn. Mit abgeschaltetem Motor wurde nur selten gesammelt, daher sind kaum Beispiele für nicht verrauschte Sprachdaten enthalten. Die protokollierten Geschwindigkeitsbereiche verteilen sich glockenförmig mit einem Maximum bei 60 bis 90 km/h. Die *„Lüftung"* war bei fast 80 % der Äußerungen auf „schwach" und die meiste restliche Zeit auf „mittel" gestellt.

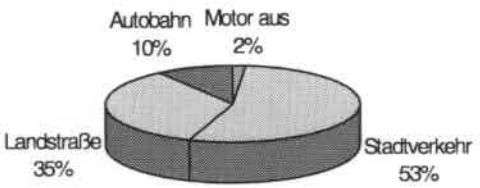


Abbildung 4.1: Anteil der Äußerungen je Fahrsituation für die freien Szenarien.

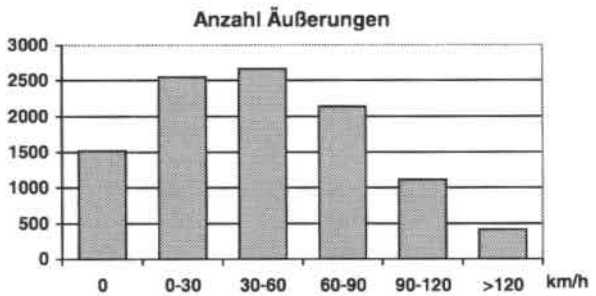


Abbildung 4.2: Anzahl der Äußerungen je Geschwindigkeitsbereich für die freien Szenarien.

Von den 43 Sprechern, deren Äußerungen in 3 verschiedenen Fahrzeugtypen (Mittelklassewagen) gesammelt wurde, sind 10 als Testsprecher bestimmt worden. Da diese Testsprecher hauptsächlich Sprache für das festgelegte Szenario (s.u.) spendeten, bleiben zum Trainieren noch knapp über 9.000 Äußerungen von 33 Sprechern. Die restlichen Äußerungen der 10 Testsprecher innerhalb des offenen Szenarios wurden von der Trainingsmenge ausgeschlossen.

Festgelegte Szenarien

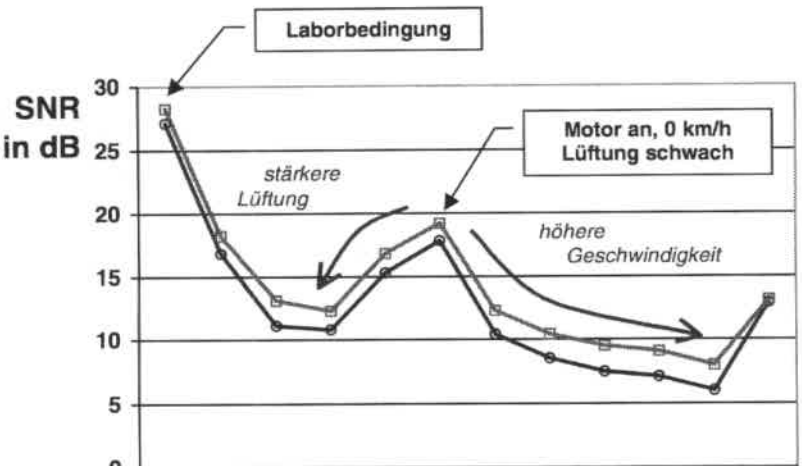
Die *festgelegten Szenarien* sind zur Evaluation von Spracherkennern gedacht. Es handelt sich dabei um Navigationsszenarien bei dem jeder Sprecher bei mehreren, fest definierten Umgebungssituationen jeweils die gleichen Äußerungen spricht. Für 12 unterschiedliche *Kategorien* wurden dafür die Umgebungsbedingungen festgelegt (im Gegensatz zu den nicht auf diese Weise beschränkten freien Szenarien oben). Dies soll ermöglichen, die Leistungsfähigkeit der Erkenner in verschiedenen Situationen zu beurteilen (siehe **Tabelle 4.2**).

Tabelle 4.2: Definition der Kategorien 01 bis 12 der festgelegten Szenarien.

Nr.	Bedingungen: Geschwindigkeit in km/h	Lüftung	Besonderheit
01	0	Aus	**Motor aus**
02	0	schwach	**Blinker**
03	25	schwach	**Fenster auf**
04	0	**stark**	
05	0	**mittel**	
06	**0**	**schwach**	
07	**25**	schwach	
08	**50**	schwach	
09	**75**	schwach	
10	**100**	schwach	
11	**125**	schwach	
12	**ca. 20 bis 50**	schwach	**beschleunigen**

Insgesamt ergeben sich für die 10 Testsprecher, die für jede Kategorie jeweils 30 Anfragen stellten, 3.600 Äußerungen. Diese wurden wie auch die anderen Daten gleichzeitig über die oben erwähnten Mikrofone aufgenommen. Die 300 Anfragen aller Sprecher pro Kategorie sind vom Umfang her ein Testset üblicher Größe. Sie lassen sich gut mit den wortgleichen, aber sicherlich etwas anders gesprochenen Äußerungen (auch bedingt durch den Lombard-Effekt!) der anderen Kategorien vergleichen.

Abbildung 4.3 erlaubt eine Abschätzung der Schwierigkeiten bei der Erkennung. Sie zeigt den Verlauf des mittleren modifizierten Mean-SNR mit zwei unterschiedlichen Sprach-Pause-Detektoren. Auch wenn die Werte konsistent voneinander abweichen, bleibt der Trend gleich. Erwartungsgemäß besitzt die laborähnliche Kategorie 01 den größten SNR-Wert. Nach Einschalten des Motors und der Lüftung (Kategorie 06) ergibt sich ein drastischer Verlust von beinahe 10 dB, der sich bei Erhöhung der Geschwindigkeit (Kategorie 06 bis 11) ebenfalls weiter erhöht. Eine Erhöhung der Lüftereinstellung (Kategorien 06, 05, 04) reduziert ebenfalls den SNR, aber nur bis auf ein Niveau, das etwa einer Geschwindigkeit von 25 km/h entspricht. Es sei darauf hingewiesen, dass der tatsächliche SNR (und erst recht der gemessene) nicht direkt über die Zunahme des Hintergrundrauschens Auskunft gibt. Bei Zunahme der Umgebungsgeräusche erhöhen viele Sprecher sozusagen ihre Signalenergie und gleichen damit das verstärkte Rauschen etwas aus. Auf diesen Punkt wird in den späteren Analysen nochmals eingegangen.



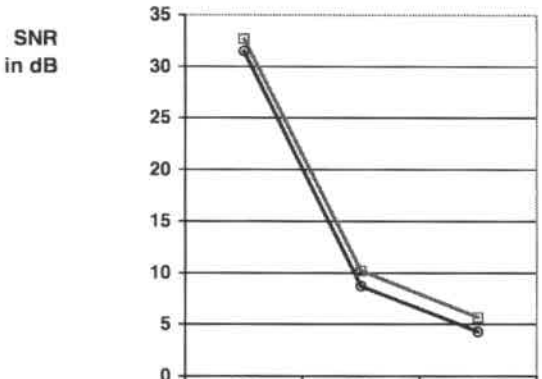
	01	02	03	04	05	06	07	08	09	10	11	12
Besonderheit	aus	Blinker	Fenster	Lstark	Lmittel	0km/h	25km/h	50km/h	75km/h	100km/h	125km/h	beschl.
ModSNR1	27,1	16,8	11,1	10,8	15,3	17,8	10,4	8,5	7,5	7,1	6	12,9
ModSNR2	28,3	18,2	13,1	12,3	16,8	19,2	12,3	10,4	9,5	9,1	8	13,1

Abbildung 4.3: Modifizierter Mean-SNR mit zwei verschiedenen Sprach-Pause-Detektoren für die 12 Testkategorien (Kanal 1, AKG).

Geräuschaufnahmen

Während der Sammlung wurden auch längere Geräuschaufnahmen (ohne Sprache) gemacht. Sie wurden in dieser Arbeit zur Simulation von verrauschten Trainingsdaten verwendet oder zum direkten Vergleichen mittels eines ungestörten Testsets, das einmal ohne und ein andermal mit zugemischtem Rauschen getestet wurde.

Abbildung 4.4 zeigt wiederum den mittleren modifizierten SNR, diesmal für das Nahbesprechungsmikrofon und zugemischtes Rauschen (Mischfaktor 0,0 bis 1,0). Obwohl Rauschen und Sprachsignal auch getrennt vorlagen, wurde die SNR-Bestimmung genau wie für **Abbildung 4.3** mit Sprach-Pause-Detektoren durchgeführt. Für das Fahrgeräusch, das bei einer Geschwindigkeit von 100 km/h aufgenommen wurde, ergeben sich deutlich schlechtere Werte als bei der vergleichbaren Kategorie 10 der **Abbildung 4.3** (ungefähr 5 dB statt etwa 8 dB). Es bestätigt sich also schon durch einfache SNR-Bestimmung ein gewisser Unterschied im Vergleich zwischen realen und simulierten Daten.



	0,0	0,5	1,0
Katgorie, Kanal	01, Kanal 2	01, Kanal 2	01, Kanal 2
ModSNR1	31,5	8,7	4,3
ModSNR2	32,7	10,2	5,7

Abbildung 4.4: Modifizierter Mean-SNR mit zwei verschiedenen Sprach-Pause-Detektoren für die ungestörte Kategorie 01 (Kanal 2, Sennheiser) mit zugemischtem Fahrgeräusch (100km/h).

4.1.3 Zusammenfassung

Tabelle 4.3 gibt nochmals einen Überblick der drei wichtigsten Sprachdatenbestände, die in dieser Arbeit benutzt wurden. Daneben fanden noch reine Geräuschdaten und in Sonderfällen Testsets mit Kommandowörtern und Straßennamen Verwendung. Die Fahrgeräusche wurden sowohl zum Trainieren (künstlich verrauschte GSST-Daten) als auch für Tests (künstlich verrauschte Kategorie 01) verwendet, jedoch nie beide Fälle in Kombination.

Tabelle 4.3: Sprachdaten zum Trainieren und Testen.

Sprachdaten	Zweck (Gesamtdauer):	Anzahl:
GSST (Labor) Einkanalig (Nahbesprechungs-mikrofon)	Training (30 h)	14.400 Äußerungen 1.750 Sprecher
Autosprachdaten Zweikanalig (Nahbesprechungs- und Raummikrofon)	Training (12 h)	9.000 Äußerungen 33 Sprecher
	Test (festgelegtes Szenario)	12 Kategorien zu je: 300 Äußerungen 10 Sprecher

4.2 Basissystem

In diesem Abschnitt wird das **Basissystem** beschrieben. Gemeint ist damit ein spezieller Erkenner für deutsche Spontansprache, der als Ausgangspunkt der weiter unten vorgestellten Experimente diente. Seine „Konstellation", also die Art der Modellierung, Wahl der freien Parameter, Trainingsverfahren usw., wird hier gegenüber den allgemeinen Grundlagen im Kapitel 2 konkretisiert. Sie entspricht dem neuesten Stand der Technik von kontinuierlichen Erkennern für Spontansprache und ist das Ergebnis langjähriger Zusammenarbeit der Mitarbeiter der *Interactive Systems Laboratories (ISL)*[12]. Wörterbuch und Sprachmodell sowie die damit zusammenhängenden Parametereinstellungen mussten für die Erkennung von spontansprachlichen Navigationsanfragen angepasst werden.

[12] ISL steht für die beiden Forschungsgruppen unter Prof. A. Waibel an der Universität Karlsruhe und an der Carnegie Mellon University in Pittsburgh (USA).

Entwicklungswerkzeug

Der Basiserkenner sowie alle anderen für diese Arbeit verwendeten Systeme basieren auf dem **Janus-RecognitionToolkit (JRTk)**. Dieses wurde an der Universität Karlsruhe, später auch an der Carnegie Mellon University Pittsburgh, entwickelt. Das Programmpaket enthält die heute gängigen Standardverfahren eines kontinuierlichen Spracherkenners und wurde darüber hinaus im Laufe der letzten Jahre um zahlreiche Ergänzungen bereichert (vom Autor selbst stammen die meisten Module der Merkmalsextraktion sowie einige andere Teile). Die einzelnen Komponenten eines Spracherkenners werden dabei als Objekte betrachtet, die über eine Skriptsprache erzeugt, konfiguriert und miteinander verbunden werden können. Über Methodenaufrufe werden die Abläufe zum Trainieren und Erkennen gesteuert oder Inhalte ausgegeben. Dieses Konzept ermöglicht eine hohe Flexibilität, so dass für neue Experimente oft gar keine Codeänderungen nötig sind. Durch sehr gutes Abschneiden bei Evaluationstests, sowohl innerhalb des deutschen Projektes VerbMobil (s.o.), wie auch bei der amerikanischen (D)ARPA[13], konnte JRTk seine Leistungsfähigkeit unter Beweis stellen. Inzwischen wird es auch von anderen Spracherkennungsgruppen in Ländern wie Japan, Korea oder Frankreich eingesetzt.

Merkmalsextraktion

Abbildung 4.5 zeigt die im Basiserkenner verwendete Merkmalsextraktion. Zugrunde liegt ein digitalisiertes Zeitsignal mit 16 kHz Abtastrate und 16 bit Quantisierung. Die einzelnen Schritte können über Methodenaufrufe eines Objektes definiert werden, das die Merkmale verwaltet. So können auch leicht alternative Vorgehensweisen verwendet werden. Die Zahlen am linken Rand zeigen die Anzahl der Koeffizienten i für die jeweiligen Mekmalsvektoren. Nach der Entnahme eines Analyserahmens k mit einem Hammingfenster wird durch die FFT (engl.: *Fast Fourier Transform*), einer schnellen Implementierung der diskreten Fouriertransformation (DFT), eine Kurzzeit-Spektralanalyse durchgeführt. Das Leistungsspektrum wird berechnet und durch eine *Mel*-Filterbank zu 30 Bändern zusammengefasst. Nach der Logarithmierung wird mit der inversen DFT das Cepstrum berechnet und auf 13 Koeffizienten beschränkt („*Liftering*", Ausblenden der oberen Cepstralkoeffizienten). Die Transformation in den Cepstralbereich ist für solche Fälle eine einfache Matrixmultiplikation. Durch die Schritte „*Mel*-Filterbank" und „*Liftering*" wird die Einhüllende des Spektrums in einer kompakten Form extrahiert. Diese statische Information wird durch die dynamischen Merkmale $\Delta \mathbf{x}^c[k]$ und $\Delta\Delta \mathbf{x}^c[k]$ mit Kontextinformation erweitert. Zuvor wird jedoch noch eine Kanalkompensation mithilfe einer sprachbasierten Mittelwertsubtraktion durchgeführt (auf diese wird in Abschnitt 8.5 noch genauer eingegangen). Als letzter Schritt wird mit der Linearen Diskriminanzanalyse (LDA) der Merkmalsvektor dekorreliert und in seiner Dimension auf 32 Merkmale reduziert.

---

[13] (D)ARPA: (Defense) Advanced Research Projects Agency. Die Projekte dieser Behörde trugen wesentlich zur Entwicklung der Spracherkennung über die letzten Jahrzehnte bei. Bei den regelmäßig stattfindenden Evaluationen werden die Forschungsarbeiten international führender Spracherkennungsgruppen miteinander verglichen.

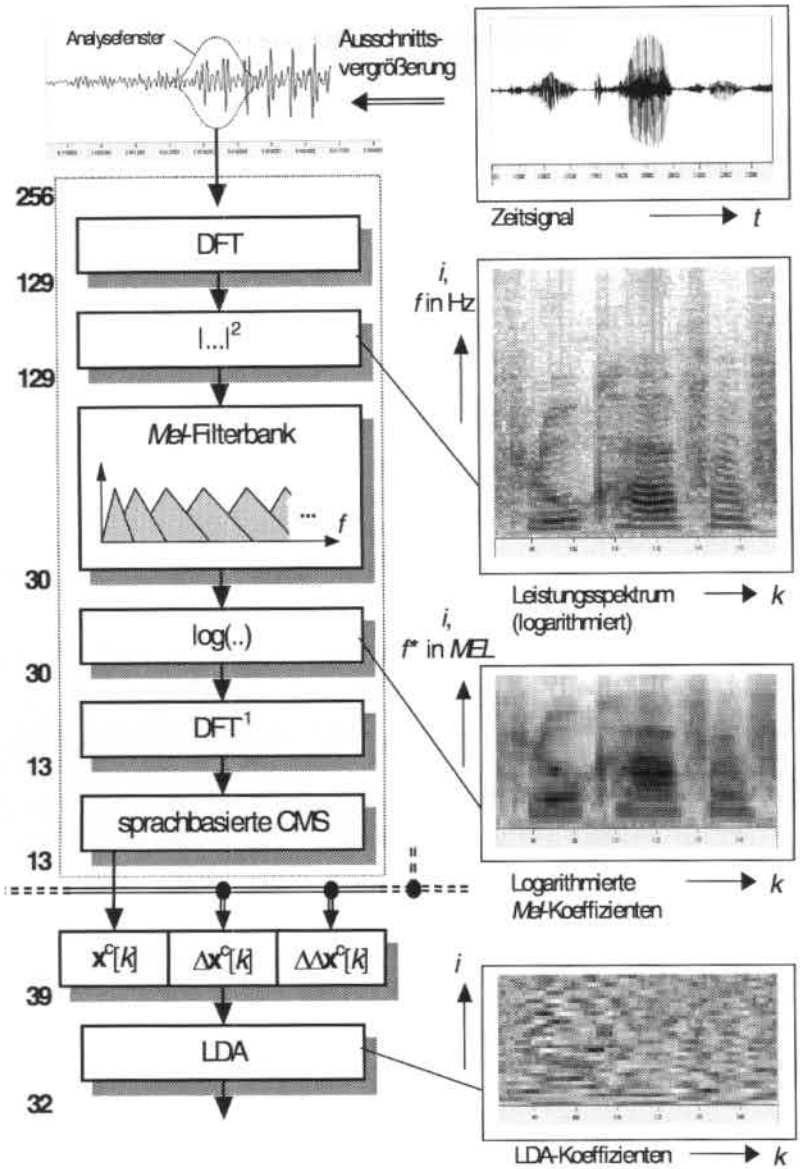


Abbildung 4.5: Merkmalsextraktion des Basiserkenners.

Akustisches Modell

Das HMM-basierte akustische Modell hält für den 32-dimensionalen Merkmalsraum Dichteschätzungen bereit, die jeden der Zustände des HMM mit einer Mixtur aus 32 Gaußverteilungen beschreiben. Die einzelnen Gaußverteilungen besitzten diagonale Kovarianzmatrizen. Da kontinuierliche Dichten und nicht Codebücher mit Vektorquantisierung benutzt werden, handelt es sich um ein kontinuierliches HMM (im Gegensatz zum diskreten HMM).

Den **Zuständen** liegen phonetische Einheiten zu Grunde. Um den Verlauf eines Phonems zu beschreiben, wird Beginn-, Mittel- und Endphase unterschieden (später ***Phonemteile*** genannt). Der phonetische Kontext wird bei dem jeweiligen Modell berücksichtigt, um Koartikulationseffekte zu erfassen. Nach links und rechts werden jeweils bis zu zwei Phoneme betrachtet. Dies ergibt also mitsamt der eigentlichen Phonemidentität fünf Einflüsse, die die Charakteristik des Lautes bestimmen. Diese spezielle Form **kontextabhängiger Phoneme** wird daher ***Quintphone*** genannt (in den meisten Erkennern sind hingegen nur *Triphone* üblich). Liegen die benachbarten Phoneme innerhalb desselben Wortes, werden sie alle als Kontext betrachtet, wortübergreifend aber maximal nur eines. Damit unterscheidet sich ein Phonemteil, beispielsweise der Beginn eines bestimmten Vokals, durch die

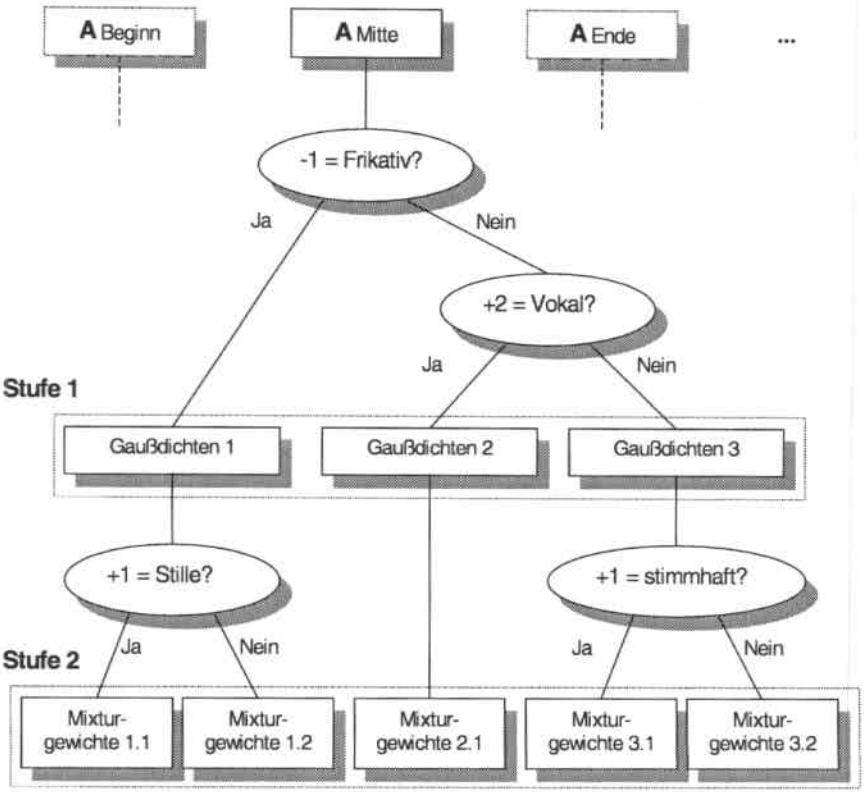
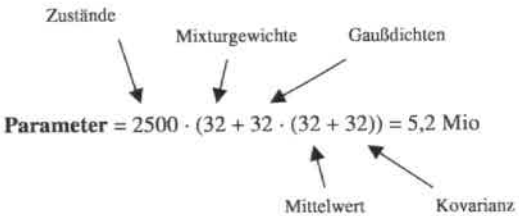


Abbildung 4.6: Zweistufige divisive Ballung mit Entscheidungsbaum.

benachbarten vier Phoneme von Vorkommen desselben Vokalteils in anderem Kontext.

Die große Zahl der sich so ergebenden Einheiten wurde durch ein **zweistufiges Ballungsverfahren** reduziert (siehe **Abbildung 4.6**). Dieses benutzt ein Entropie-Maß, mit dem phonetisch basierte Fragen zur divisiven Ballung ausgesucht werden. So entsteht ein Entscheidungsbaum, der die Menge aller Kontexte eines Phonemteils immer weiter aufspaltet. An den Blättern dieses Baumes befinden sich dann jeweils Gruppen mit ähnlichen Kontexten. Der Vorteil, der sich aus der zweistufigen Anwendung ergibt, ist der Freiheitsgrad zwischen kontinuierlichem und semi-kontinuierlichem HMM. Semi-kontinuierlich bedeutet, dass Zustände zwar eigene Mixturgewichte besitzen, die zugehörigen Dichtefunktionen aber mit den anderen Zuständen teilen. Nach anfänglicher Ballung zu 10.000 unterschiedlich modellierten Zuständen (Stufe 2) und 2.500 Sets von Gaußdichten (Stufe 1) wurde die Zahl der Zustandsmodelle ebenfalls auf 2.500 reduziert (womit ein rein kontinuierliches HMM vorliegt). Bei nur sehr geringen Verlusten in der Erkennungsrate wurde damit die Zahl der Parameter verkleinert sowie die Erkennungsgeschwindigkeit erhöht.

Durch Training zustandsabhängiger Übergangswahrscheinlichkeiten ließen sich bisher keine Gewinne erzielen, so wurden diese einmalig für alle Zustände gemeinsam bestimmt. Die Gesamtzahl der Parameter für das akustische Modell ergibt sich daher zu:



Dazu kommen noch Parameter für **beschleunigende Maßnahmen**. Mit einem zusätzlichem, kleineren akustischen Modell für kontext**u**n**a**bhängige Phoneme werden bei der Erkennung vorausschauend Bewertungen berechnet und so der Suchraum eingegrenzt (engl.: ***phoneme look-ahead***). Es handelt sich hier um etwa 150 unterschiedlich modellierte Zustände mit weniger Gaußdichten pro Mixtur und ebenfalls weniger Merkmalen. Bei der Auswertung des genaueren, kontextabhängigen Modells werden außerdem nicht alle Gaußdichten berücksichtigt, sondern nur diejenige, deren Mittelwertvektor den kleinsten Abstand zum aktuellen Merkmalsvektor besitzt. Um die entsprechende Gaußdichte schneller zu finden, wird das **BBI-Verfahren** (engl.: *bucket box intersection*) [Fritsch und Rogina 1996] eingesetzt, das ebenfalls nochmals eigene Parameter benötigt. Diese Verfahren zur Beschleunigung sind ausführlich in [Woszczyna 1998] beschrieben und werden bei der Verwendung des Erkenners im Demonstratorsystem eingesetzt. Da hier ein Kompromiss zwischen Erkennungsgenauigkeit und Geschwindigkeit eingegangen wird, wurden diese Maßnahmen bei den weiter unten vorgestellten Experimenten nicht verwendet. Zwar hätten diese die Durchführung der Versuche beschleunigt, die Ergebnisse aber vor allem zum Nachteil des Basissystems verfälscht.

Phoneminventar

Das verwendete **Phoneminventar** zur Beschreibung der Wortaussprachen orientiert sich überwiegend am international standardisierten IPA (Internationales Phonetisches Alphabet). Neben regulären Phonemen wurden spezielle Laute für Geräusche menschlichen und nichtmenschlichen Ursprungs eingeführt. Diese beschreiben „Geräuschwörter", die neben regulären Wörtern im Aussprachewörterbuch aufgenommen wurden. Die Gesamtzahl aller Aussprachеeinheiten inklusive Pause beträgt 61.

Für diese Einheiten wurden **phonetische Eigenschaften** (z.B. „Vokal" oder „stimmhaft") festgeschrieben. Mithilfe dieser Eigenschaften werden die oben erwähnten Fragen für die Ballung definiert. Mit der Frage „Ist das zweite Vorgänger-Phonem stimmhaft?" wird die betrachtete Menge von kontextabhängigen Phonemen in zwei Gruppen geteilt. Denkbar wäre auch eine Dreiteilung (ternärer Baum), falls man eine Gruppe unterscheidet, für die diese Frage nicht beantwortbar ist. Diese Gruppe wurde jedoch wegen einiger Nachteile dieser Aufteilung (siehe [Rogina 1998]) mit der „nein"-Gruppe zusammengelegt.

Aussprachewörterbuch

Um akustische Modelle für die Phoneme oder davon abgeleitete Einheiten zu trainieren, braucht man verschriftete Trainingsdaten. Diese wurden weiter oben vorgestellt und enthalten im Falle der GSST-Daten und einem Teil der im Auto gesammelten Daten auch die bereits erwähnten Geräuschwörter. Aus den Wortfolgen wird mithilfe eines **Aussprachewörterbuches** und den HMM-Bausteinen der einzelnen Phoneme ein Satz-HMM mit optionalen Pausenzuständen aufgebaut. Für alternative Aussprachen werden parallel laufende Zustandsketten angelegt.

Grundsätzlich ist also für jedes Wort der Trainingsmenge ein Eintrag mit der Aussprache in Form einer oder mehrerer Phonemfolgen nötig. Dazu stand ein großes Aussprachewörterbuch mit über 300.000 Hauptformen und Alternativaussprachen zur Verfügung. Darin wurden zum Teil auch abgebrochene Wörter aufgenommen, wie sie zuweilen in spontaner Sprache vorkommen und auch verschriftet wurden. Bei den Navigationsanfragen kamen sehr viele Namen (zum Beispiel von Straßen und Orten) vor, die in diesem Wörterbuch nicht vertreten waren. Teilweise konnten diese Namen aus existierenden Einträgen zusammengesetzt werden. Für andere Fälle wurde zunächst mit einem Graphem-zu-Phonem-Konverter („Schrift" zu „Laut", wie er auch bei der Sprachsynthese verwendet wird) eine erste phonetische Umschreibung erzeugt und danach manuell korrigiert. Die manuelle Korrektur war notwendig, da die üblichen Regeln zur Aussprachegenerierung bei Namen oft versagen. Zudem treten auch im Szenario (im Raum Karlsruhe) englische Straßennamen auf (z.B. *„Indiana Lane"*, *„New-Jersey-Street"*). Für diese wurde eine in etwa zutreffende Aussprache mittels des deutschen Phoneminventars gewählt (siehe zu diesem Thema auch [Trancoso *et al.* 1999]).

Seltene, falsch geschriebene oder abgebrochene Wörter, die nicht im Aussprachewörterbuch zu finden waren, wurden auf ein *„mumble word"* (Gemurmel) mit einer Folge aus generischen Lauten abgebildet. So können einerseits die übrigen Satzteile zum Trainieren der spezifischen Phoneme verwendet werden und außerdem mit dem *„mumble word"* ein einfaches Modell für unbekannte Wörter erstellt werden.

Die verschrifteten Geräusche wurden ebenfalls abgebildet. Da die vorgesehene Anwendung sich nicht auf die Labor-/Büroumgebung beschränkt, wurden die nicht von Sprechern stammenden typischen Bürogeräusche zu einem einheitlichen „Müll"-Wort zusammengefasst. Verwendet wurden die folgenden Geräuschwörter:

+AEH+, +AEHM+, +ATMEN+, +HM+, +MUELL+

Abbildung 4.7 zeigt einige Einträge des Aussprachewörterbuches und die dazu gehörigen Laute. Die Darstellung wurde hier gegenüber dem tatsächlich verwendeten Format etwas vereinfacht. Die Geräuschwörter (oben) bestehen jeweils aus nur einem Geräuschlaut, die jedoch aus mehreren unterschiedlich modellierten Zuständen bestehen. Für das Wort „komme" finden wir eine Aussprachevariante. Das jeweils erste und letzte Phonem einer Folge ist mit einer Markierung „WB" für die Wortgrenze (engl.: *word boundary*) versehen. Bei der Ballung kontextabhängiger Phoneme können so, neben Fragen über phonetische Eigenschaften, auch Fragen nach bestimmten Markierungen gestellt werden.

+AEH+	$+hEH_{WB}$
+AEHM+	$+hEM_{WB}$
+ATMEN+	$+hBR_{WB}$
⋮	⋮
komme	$K_{WB}\ O\ M\ E2_{WB}$
komme(komm')	$K_{WB}\ O\ M_{WB}$
kommen	$K_{WB}\ O\ M\ E2\ N_{WB}$
kommt	$K_{WB}\ O\ M\ T_{WB}$
⋮	⋮

Abbildung 4.7: Aussprachewörterbuch mit Geräuschwörtern, Varianten und Markierungen.

Training

Der gesamte Trainingsprozess, von einer gegebenen Sprachdatenbank und zufällig initialisierten akustischen Modellen bis hin zu einem fertigen kontinuierlichen Erkenner, ist sehr langwierig. Zur Verkürzung werden die Sprachdaten daher meist mit einem bereits bestehenden Erkenner phonetisch etikettiert. Im vorliegenden Fall bedeutet dies, dass jedem Sprachrahmen der Trainingsmenge mithilfe des Viterbi-Algorithmus ein Zustand (Phonemteil) zugeordnet wird.

Zunächst wurde mit einem kontextunabhängigen System begonnen, das mit dem schon angesprochenen zweistufigen Ballungsverfahren in ein kontextabhängiges, kontinuierliches HMM-System überführt wurde. Mit diesem wurden dann sprecheradaptierte Viterbi-Zuordnungen erstellt. Sie dienten als Ausgangspunkt des Standardtrainings, sowohl für das Basissystem als auch für alle anderen hier benutzten Vergleichssysteme. Hierfür müssen noch die beiden Begriffe *„sprecheradaptiert"* und *„Standardtraining"* geklärt werden:

Bei der **sprecheradaptiven Viterbi-Zuordnung** wird das akustische Modell mit allen Äußerungen eines Trainingssprechers mit der MLLR (siehe Abschnitt 6.4.1, S. 96) basierend auf einer ersten Viterbi-Zuordnung adaptiert. Der Erkenner wird dadurch auf diesen Sprecher spezialisiert. Nun werden die Viterbi-Zuordnungen erneut berechnet und für das folgende Standardtraining abgespeichert. Für den nächsten Sprecher wird wieder mit dem ursprünglichen Modell begonnen, mit seinen Äußerungen adaptiert, die Zuordnungen für diese berechnet und so weiter. Dadurch ergeben sich ähnlich gute Zuordnungen wie für ein sprecherabhängiges System. Von diesen weiß man, das sie wesentlich geringere Fehlerraten liefern als vergleichbare sprecher**un**abhängige.

Das **Standardtraining** beruht nur auf gegebenen Viterbi-Zuordnungen und die damit zusammenhängenden Grundlagen. Gemeint ist hiermit die Aufteilung der entsprechenden Zeitsignale in Sprachrahmen und die Folge von phonetischen Einheiten, die dieser Aufnahme zugeordnet wurde. Das bedeutet, nur die Schrittweite der Merkmalsextraktion und die Umschrift sowie das Aussprachewörterbuch müssen beibehalten werden. Die Parameter der Modelle, wie Anzahl der Gaußdichten, Merkmale usw., können verändert werden. Das Standardtraining besteht aus den folgenden drei Schritten, bei denen die gegebene Zuordnung festgehalten wird:

1. Bestimmung einer Transformationsmatrix zur Redundanzreduktion (LDA, siehe Abschnitt 5.4.2).
2. Initialisierung der akustischen Modelle auf einem Teil der Trainingsmenge mit dem „*Neural-Gas*"-Verfahren [Martinez *et al.* 1993], einer Verallgemeinerung des Basic-Isodata [Duda und Hart 1973].
3. 4 Trainingsiterationen über die gesamte Trainingsmenge entsprechend dem EM-Algorithmus (Abschnitt 2.4). Hierbei werden jeweils die Mixturgewichte sowie die Parameter der Gaußdichten neu geschätzt (Abschnitt 3.3.4).

Sprachmodell und Vokabular

Zur Berechnung der Wortfolgestatistiken für das (linguistische) Sprachmodell wird möglichst viel Textmaterial verwendet, um aussagekräftige Werte zu bekommen. Für Diktieranwendungen werden daher zum Beispiel große Mengen Zeitungsartikel benutzt, von denen man sich eine gewisse Ähnlichkeit zu den diktierten Texten erhofft. Für spontansprachliche Anfragen ist man dagegen auf die Verschriftungen einer entsprechenden Datensammlungen angewiesen.

Für das Sprachmodell des Basiserkenners kamen also die Texte der GSST-Trainingsmenge in Frage, die sich thematisch allerdings nur teilweise mit der hier angestrebten Anwendung überdecken. Ein Versuch entsprechende Texte mithilfe der Grammatiken der Sprachverarbeitungskomponente (engl. kurz „*parser*" genannt) zu erzeugen, lieferte nur unbefriedigende Ergebnisse, da eine starke Übergenerierung stattfand (mit sehr wortreichen Anfragen). Eine Lösung dafür hätte darin bestanden, die *Parser*-Regeln mit Wahrscheinlichkeiten zu belegen. Diese müssten aber wiederum anhand realistischer Daten geschätzt werden, also Verschriftungen, die ja gerade fehlen. So wurden etliche hundert realer Anfragen mithilfe eines künstlichen Szenarios und später mit dem Demonstratorsystem verschriftet. Parallel

dazu fand auch eine Sammlung von schriftlichen Anfragen über das Internet statt, die aber leider keine brauchbaren Daten lieferte. Ergänzt wurden die knappen Quellen durch typische Anfragen, die halbautomatisch mit Füllwörtern oder in Satzteilen variiert wurden. Insgesamt standen so knapp 20.000 Sätze zur Verfügung, die bei der Berechnung des Sprachmodells je nach Nähe zum Anwendungsszenario gewichtet wurden.

Bei der im Vergleich zu Textdaten für Diktieranwendungen relativ kleinen Menge spielt die Akustik eine größere Rolle, was die Anfälligkeit gegenüber Schwankungen durch Umgebungswechsel groß macht. Das Sprachmodell kann zum Beispiel bei der Vorhersage von einem aus insgesamt über 1.700 Karlsruher Straßennamen keine Hilfe sein, da diese als gleich wahrscheinlich vorausgesetzt wurden (s.u.). Kommt an einer Stelle in der Anfrage ein Name vor, hängt die Entscheidung nur von der akustischen Bewertung ab. Um den Einfluss des Sprachmodells bei den Versuchen nicht überzubewerten, wurde bei der Erkennung (s.u.) der dritte Durchgang, der die vollen Trigramme verwendet, nicht ausgewertet.

Nicht alle Straßennamen, die in der Anwendung vorkommen können, treten bereits in den gesammelten Texten auf. Umgekehrt sind andere Namen, eventuell Straßennamen aus anderen Städten, für die augenblickliche Anwendung nicht interessant. Daher wurden insbesondere Namen auf Klassensymbole abgebildet und als solche im Sprachmodell verwendet. Eine Klassenbildung eines solchen **klassenbasierten Sprachmodells** kann auch automatisch erstellt werden, was aber hier aufgrund der klaren Gegebenheiten nicht nötig war. Als Klassen wurden beispielsweise verwendet:

- Eigenname
- POI (engl.: *point of interest*, z.B. Sehenswürdigkeiten)
- Stadt
- Stadtteil
- Straße
- Zahl (für Hausnummern)
- Kommandowort

Für die jeweilige Anwendung werden die Klassen mit neuen Elementen aus vorbereiteten Listen besetzt. Für die hier beschriebenen Experimente und somit auch für den Basiserkenner ist das Szenario die Stadt Karlsruhe. Die bedingten Wahrscheinlichkeiten der Klassenelemente wurde auf konstante Werte gesetzt. Man könnte zwar hier argumentieren, dass zum Beispiel größere oder längere Straßen wahrscheinlicher sind (weil vielleicht mehr Leute dort wohnen), andererseits kann man auch dagegenhalten, dass nach dem Weg zu unbekannten (kleineren) Straßen häufiger gefragt werden muss.

Eine der oben aufgeführten Klassen stellt **Kommandowörter** dar. Auch hier liegt der Gedanke zugrunde, die Anwendung um beliebige Einträge dieser Art zu erweitern. Auch wenn es sich um einen Erkenner für spontansprachliche Anfragen handelt, enthält er gewissermaßen als Untermenge diese Option. Dabei können Kommandos jedoch in einen Kontext gestellt werden, z.B. „Ähm, Navigationssystem, bitte *Karte_vergrößern*!", wobei die zusammengezogenen Wörter als Kommando betrachtet werden können.

Das **Vokabular**, also der Wortschatz während der Erkennung, ergibt sich einerseits aus den ausgewählten Klassenelementen und andererseits aus den Wörtern der gesammelten Texte (ohne die Wörter, die in eine der Klassen gehören). Es wurden allerdings nur mehrfach

in den Texten vorkommende Wörter verwendet, da sonst das Vokabular durch Einzelfälle unnötig groß wird. So tauchte zum Beispiel der „Schwanensee" einmalig in den Textdaten auf, da ein Sprecher sich nach dem Weg ins Staatstheater erkundigte und dabei auch den Namen der Vorstellung nannte). Es ergaben sich so ca. 1.000 Basiswörter zur Formulierung der Anfragen, die mit den über 1.700 Straßennamen und den anderen Klassenlisten ein knapp 3.000 Wörter umfassendes Vokabular bilden. Die Erstellung eines entsprechenden Aussprachewörterbuches vollzog sich wie schon oben behandelt. Die OOV-Rate für die im Auto gesammelten Navigationsanfragen beträgt 3 %.

Erkennung und Auswertung

Im JRTk stehen bis zu drei aufeinander aufbauende **Suchdurchgänge** zur Verfügung. Mit dem jeweils vorangehenden Durchgang wird der Suchraum eingeschränkt, dann aber die in Frage kommenden Hypothesen über genauere Auswertungen (hauptsächlich das Sprachmodell betreffend) neu bewertet. Die Details und Unterschiede der einzelnen Schritte sind ausführlich in [Woszczyna 1998] beschrieben.

Für die Auswertungen wurden immer nur die jeweils beste Hypothese einer Äußerung betrachtet. Der gesamte Worthypothesengraph, den man alternativ dazu ausgeben kann, wurde hier nicht gebraucht. Er repräsentiert einen eingeschränkten (gut bewerteten) Teil des Suchraums aller möglichen Hypothesen. Die beste Hypothese ergibt sich als Pfad mit der besten Bewertung durch den Worthypothesengraph.

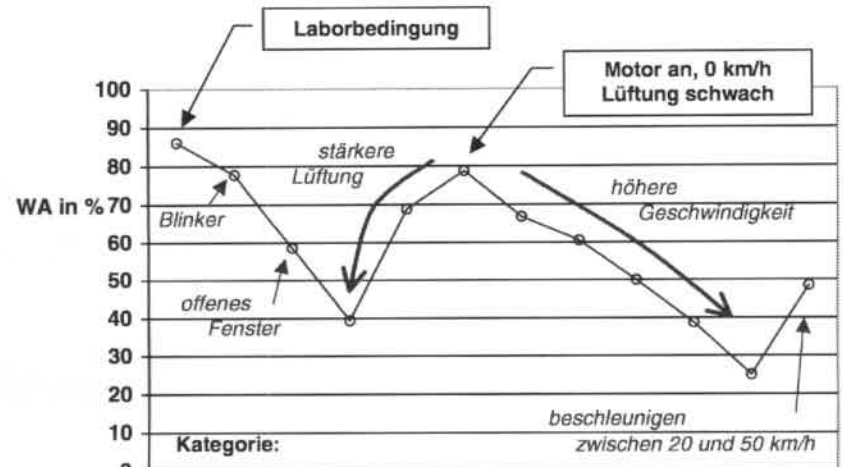
Da in dieser Arbeit ständig wechselnde Umgebungen betrachtet werden, wurde der Kontext der Erkennung auf die jeweils aktuelle Äußerung beschränkt. Das bedeutet, es stehen keine Daten aus vorangegangenen Äußerungen zur Verfügung, die Aufschluss über Sprecher oder Kanal geben könnten.

Würde man ein Absinken der Genauigkeit eines Laborerkenners auf andersartigen Daten noch erwarten, so ist man vielleicht doch von dem ebenfalls sehr starken Nachlassen der Erkennungsgeschwindigkeit überrascht. Das hängt damit zusammen wie zeitsynchrone Suchverfahren den ansonsten zu großen Suchraum beschränken. Mit einem „**Strahl**" (engl.: *beam*), der sich an dem aktuell besten Pfad orientiert, werden die noch weiter zu untersuchenden Pfade auf diejenigen eingeschränkt, die innerhalb des Strahls liegen. Mit anderen Worten, ihre momentane Bewertung weicht nur um einen bestimmten Betrag vom derzeitigen Favoriten ab. Bei gestörten Eingangsdaten ergeben sich viele gleichartige (schlechte) Bewertungen, so dass die Anzahl der „Kandidaten im Strahl" größer ist. Die Beschränkung auf eine bestimmte, kleinere Anzahl von Hypothesen führt im Extremfall zu noch schlechteren Endergebnissen, da bei gestörten Daten die lokalen Abweichungen nun mal größer sind.

Bei der Erkennungsleistung interessiert zunächst wie gut der Laborerkenner auf rauscharmen Sprachdaten abschneidet, die außerdem mit dem selben Mikrofontyp wie die Trainingsdaten aufgenommen wurden. Aus den jeweils 300 Äußerungen umfassenden Testsets aus der Autoumgebung wurde daher das entsprechende Set (Kategorie 01: Motor und Lüftung aus, Kanal 2: Sennheiser-Mikrofon) verwendet. Die Wortakkuratheit für die Hypothesen des Basiserkenners nach dem zweiten Suchdurchgang betrug 86,9 %. Eine Vergrößerung des Suchstrahls und damit des Suchraumes brachte hier keine Veränderung der

Genauigkeit, die benötigte Zeit verdoppelte sich hingegen. Selbst bei den gestörten Kategorien (s.u.) ergaben sich so nur leichte Verbesserungen bis maximal 1 % absolut, weshalb die ursprünglichen Einstellungen für alle weiteren Tests verwendet wurden.

Die Ergebnisse für das im Auto installierte Mikrofon AKG zeigt **Abbildung 4.8**. Für die Kategorie 01 (exakt wie oben, nur anderer Kanal) ergeben sich geringe Verluste von nicht mal einem Prozent absolut. Wird jedoch der Motor und die Lüftung eingeschaltet fällt die Akkuratheit schon deutlich unter 80 % (Kategorie 06). Noch schwerwiegendere Einbußen erleidet das System bei stärkeren Lüftereinstellungen und bei höherer Geschwindigkeit. Bei Tempo 100 km/h (10) zum Beispiel, ist die Akkuratheit unter 40 %, d.h. die Wortfehlerrate hat sich gegenüber Kategorie 01 mehr als vervierfacht. Für die übrigen Kategorien, wie „Blinker" (02), „offenes Fenster" (03) und „beschleunigen" (12) findet man damit verglichen Werte in einem mittleren Bereich.



	01	02	03	04	05	06	07	08	09	10	11	12
Besonderheit	aus	Blinker	Fenster	Lstark	Lmittel	0km/h	25km/h	50km/h	75km/h	100km/	125km/	beschl.
—⊖— AKG	86,1	77,7	58,6	39,4	68,9	78,7	66,9	60,5	50	38,7	25	48,5

Abbildung 4.8: Erkennungsergebnisse (Wortakkuratheit) des Laborsystems auf den einzelnen Testkategorien im Auto.

Zusätzliche Näherungen zur Beschleunigung der Erkennung (siehe oben, S. 63) hätten zu noch größeren Verlusten geführt und wurden daher nicht verwendet. Zum Beispiel verringert die Bewertung durch Selektion und Auswertung nur einer Gaußdichte pro Mixtur die Wortakkuratheit schon für den ungestörten Fall um 1 % absolut. Kommen noch BBI und *„phoneme look-aheads"* hinzu, verliert man 5 %. Bei den anderen Kategorien erhöht sich dieser zusätzliche Verlust bis auf über 15 % absolut.

Vergleicht man die dennoch bestehenden Einbußen der Erkennungsleistung mit den SNR-Werten von **Abbildung 4.3** (Seite 57), ergibt sich zwar tendenziell eine Übereinstimmung, der Zusammenhang ist aber stark von der Geräuschart und dem SNR-Bereich abhängig.

Beispielsweise nimmt die Wortakkuratheit fast linear (umgekehrt proportional) mit der Fahrzeuggeschwindigkeit ab, während die SNR-Werte eine gewisse Sättigungsphase erreichen. Bei kleineren SNR-Werten bewirken also Änderungen dieses Wertes eine stärkere Auswirkung auf die Erkennungsleistung. Die **Abbildung 4.9** zeigt den Zusammenhang für einige Kategorien (Labor: Kategorie 01, Geschwindigkeit: Kategorie 06 bis 11, Lüftung: Kategorie 06 bis 04). Es bleibt festzuhalten, dass der SNR-Wert zwar Hinweise auf die Schwierigkeiten bei der Erkennung geben kann, aber kein eindeutiger Zusammenhang mit der Erkennungsrate existiert.

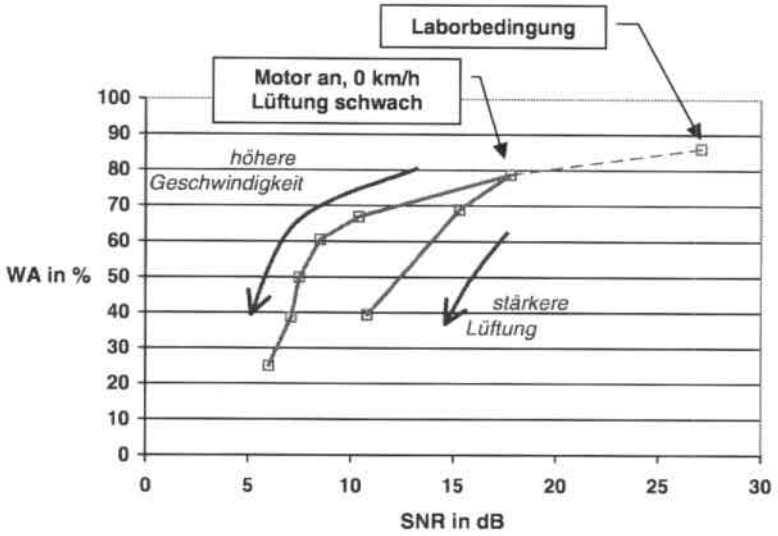


Abbildung 4.9: Die Erkennungsleistung des Basiserkenners in Abhängigkeit der gemessenen SNR-Werte (Sprach-Pause-Detektor 2).

5 Ansatzpunkte zur robusten Spracherkennung

Aufgabe der Wissenschaft ist es nicht nur,
die Erfahrung stets zu erweitern,
sondern in diese Erfahrung eine Ordnung zu bringen.
- Niels Bohr -

Kapitelübersicht: In diesem Kapitel wird das Thema robuste Spracherkennung eingeführt. Der erste Abschnitt nimmt zunächst eine Einteilung verschiedener Ansatzpunkte vor. Dort findet der Leser auch einen genaueren Überblick der noch folgenden Abschnitte und Kapitel. Einige Ansätze und dazu existierende Verfahren werden bereits in den nachfolgenden Abschnitten vorgestellt und auch teilweise experimentell ausgewertet. Die Schwerpunkte, also die Hauptbeiträge und Neuerungen dieser Arbeit, werden in eigenen, danach folgenden Kapiteln erörtert. Dabei handelt es sich um Modellanpassung und Kompensationsmaßnahmen für verschiedene Störungen.

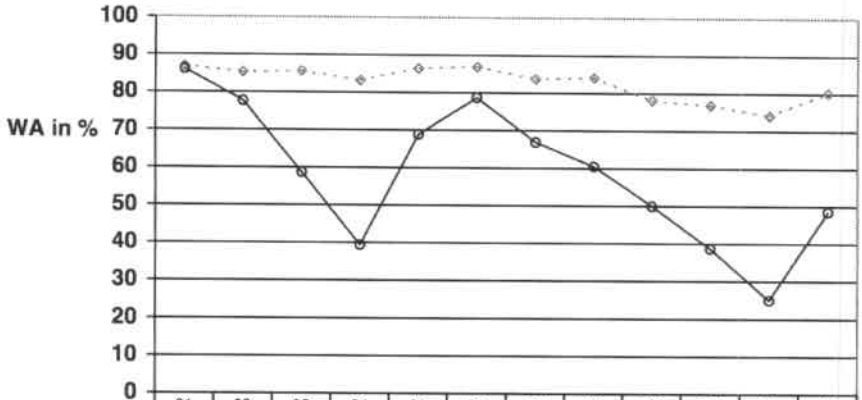
Der *experimentelle Rahmen* sei hier nochmals kurz zusammengefasst: Das Ziel dieser Arbeit ist der Einsatz der kontinuierlichen Spracherkennung in mobilen Informationssystemen. Ausgangspunkt ist das im vorigen Kapitel beschriebene **Basissystem**, das bereits auf spontansprachliche Anfragen (in Laborumgebung) optimiert wurde. Als Testszenario wird die Autoumgebung betrachtet, da hierfür durch die in Abschnitt 4.1.2 beschriebene Sammlung geeignetes Testmaterial vorliegt. Mit den 12 unterschiedlichen Testkategorien sind hinreichend verschiedene Bedingungen abgedeckt, so dass eine Übertragbarkeit auf ähnliche Szenarien erwartet werden kann. Durch die Verfügbarkeit größerer Mengen an Trainingsmaterial für diese Umgebung kann außerdem bestimmt werden, welche Ergebnisse ein Erkenner liefert, der mit großem Aufwand speziell auf diese Umgebung trainiert wurde. Dieser dient zusammen mit dem Basissystem und einem weiteren Erkenner, der mit künstlich verrauschten Trainingsdaten erstellt wurde, als Vergleichsbasis für die Experimente der folgenden Kapitel. Diese **Referenzsysteme** werden im Kapitel 6 (Modellanpassung) vorgestellt.

5.1 Ansatzpunkte

Vermeidung der Ursachen

Im Einführungskapitel dieser Arbeit wurde gezeigt mit welcher Art von Störungen ein automatisches Spracherkennungssystem im mobilen Einsatz zu tun hat. Durch wechselnde Sprecher, Übertragungskanäle und Hintergrundgeräusche wird die Varianz innerhalb der verschiedenen Sprachmuster erhöht und so die Erkennungsaufgabe erschwert. Eine der nahe-

liegenden Praktiken ist es, die Ursachen dieser Probleme zu vermeiden. Heutzutage erhältliche Diktiersysteme verfolgen diese Strategie dadurch, dass der Einsatz auf geräuscharme Sprachaufnahmen beschränkt ist und der Sprecher das System erst auf seine Stimme einlernen muss. Selbst in der Autoumgebung kann man passable Erkennungsergebnisse erreichen, wenn man zur Aufnahme ein Nahbesprechungsmikrofon verwendet und so die Sprachaufnahmen relativ rauschfrei hält (siehe **Abbildung 5.1**). Allerdings ist ein solches Mikrofon teuer und muss jedes Mal durch einen Kopfbügel oder ähnliche Halterungen nahe an den Mund des Sprechers gebracht werden. Bei mehreren Benutzern wäre eventuell für jeden einzelnen ein solches Mikrofon notwendig.



	01	02	03	04	05	06	07	08	09	10	11	12
Besonderheit	aus	Blinker	Fenster	I.stark	I.mittel	0km/h	25km/h	50km/h	75km/h	100km/	125km/	beschl.
Sennheiser	86,9	85,2	85,6	83,2	86,5	87	83,7	84,1	78	76,8	74,2	80,2
AKG	86,1	77,7	58,6	39,4	68,9	78,7	66,9	60,5	50	38,7	25	48,5

Abbildung 5.1: Erkennungsraten des Laborerkenners für Aufnahmen mit einem Nahbesprechungsmikrofon (Sennheiser) und einem fest installierten Raummikrofon (AKG) für die 12 Aufnahmebedingungen im Auto.

Mit solchen Kompromissen lässt sich die Spracherkennung zumindest vorläufig überhaupt anwenden. Zufriedenstellend ist dies aber nicht, denn von der zwischenmenschlichen Kommunikation wissen wir, dass Sprache von beliebigen Sprechern geäußert auch in gestörter Umgebung verstanden werden kann. Daher soll nun versucht werden, mit anderen Möglichkeiten auch auf gestörten Daten gute Ergebnisse zu erreichen. Wenn nicht explizit erwähnt, beruhen alle weiter unten aufgeführten Ergebnisse auf Aufnahmen, die mit einem fest im Fahrzeug installierten Raummikrofon (AKG) erstellt wurden.

Einteilung der Verfahren

Zahlreiche Forschungsarbeiten haben sich mit dem Thema „Robuste Spracherkennung" beschäftigt (einen Überblick gibt z.B. [Gong 1995]) und man findet meist eine Unterteilung in drei „klassische" Ansatzpunkte:

- Robuste Merkmale
- Anpassung der Modelle (Adaption)
- Kompensation der Störungen

Diese Ansatzpunkte werden im Folgenden kurz erläutert und es wird jeweils auf die entsprechenden Abschnitte oder Kapitel innerhalb dieser Arbeit verwiesen.

Robuste Merkmale wurden schon im Abschnitt 2.2 Merkmalsextraktion angesprochen und sind Grundvoraussetzung eines leistungsfähigen Spracherkenners: So wird zum Beispiel die meist nur als Signalträger dienende und stark sprecherabhängige Anregungsfunktion durch Filterbänke oder Liftering unterdrückt. Treten Störungen verstärkt in bestimmten Frequenzbereichen auf, könnten auch diese durch entsprechende Filter unterdrückt werden. In anderen Situationen möchte man jedoch auf diese Informationen nicht verzichten. Es ist daher schwierig für den generellen Fall robuste Merkmale zu finden. Abschnitt 5.4 widmet sich diesem Thema.

Im Kapitel 6 werden verschiedene Methoden der **Modellanpassung** für veränderte Umgebungen vorgestellt und verglichen. Aus eigener Erfahrung wissen wir, dass ein ungewohnter Dialekt oder Sprechstil uns zunächst Schwierigkeiten bereitet, wir aber nach einer Gewöhnungsphase damit besser zurecht kommen. Auch ein Spracherkenner in neuer oder wechselnder Umgebung hat zunächst ein anderes akustisches Modell der Sprache, das sich aber mit neuen Daten anpassen lässt. Das oben angesprochene „Einlernen" eines kommerziellen Diktiererkenners ist eine solche Adaption der akustischen Modelle. Mit den geäußerten Sprachdaten werden die vorhandenen Modelle (die zuvor mit Aufnahmen sehr vieler Sprecher trainiert wurden) auf den aktuellen Sprecher und die Aufnahmeumgebung durch eine Transformation angepasst.

Spricht man von „**Kompensation der Störungen**" oder „Störreduktion" ist damit meist eine aktive Unterdrückung der Störeinflüsse gemeint (auch wenn die Grenze zu den „robusten Merkmalen" fließend ist). Voraussetzung ist, dass der Einfluss der Störung bekannt ist und sie durch ein geeignetes Verfahren geschätzt werden kann. Dazu verwendet man ein Modell der Umgebung mit dem der Einfluss der Störung mathematisch beschrieben werden kann (nicht zu verwechseln mit dem akustischen Modell des Erkenners). Da diese Umgebungsmodelle zum Teil auch als Grundlage für Adaptionsverfahren dienen, ergeben sich auch zu diesen Gemeinsamkeiten und Überschneidungen. Die Störkompensation ist eine der wichtigsten Maßnahmen der robusten Spracherkennung und wird daher ausführlich in Kapitel 7 (Sprechernormierung) und Kapitel 8 (Kanal- und Rauschkompensation) behandelt.

Gemeinsam ist allen Ansätzen, dass neben Annahmen über das Sprachsignal auch ein Modell der Störung zugrunde liegt. Basierend auf diesen Modellvorstellungen versucht man:

- bei den *robusten Merkmalen* **Parameter des Sprachsignals** zu bestimmen, die durch die Störung wenig beeinflusst werden
- bei der *Adaption* eine anpassende **Transformation des Erkennermodells** zu definieren und deren **Parameter** zu finden
- bei der *Kompensation* eine **Transformation des Sprachsignals** zu definieren und deren **Parameter** zu finden

Neben diesen „klassischen Ansätzen" kann man noch zwei weitere Punkte aufführen, die ein System mit Spracheingabe robuster machen:

- Ausreißerbehandlung
- Robuste Verarbeitung

Da diese eher die Symptome statt der Ursachen beseitigen und auch zum Teil in den Basiserkenner beziehungsweise die anschließenden Komponenten eingearbeitet wurden, werden sie zusammen mit den robusten Merkmalen in den nächsten Abschnitten lediglich angerissen. Die wichtigsten Ansatzpunkte zur robusten Erkennung, nämlich die Störkompensation und die Anpassung der Modelle, werden in den folgenden Kapiteln erörtert und ausgebaut.

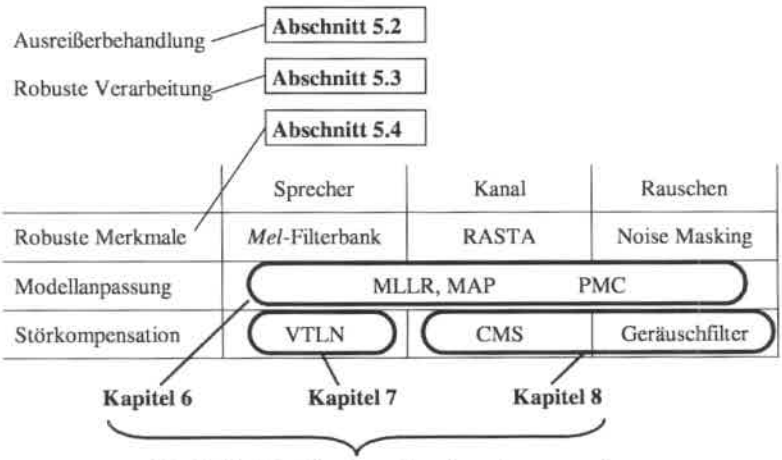


Abbildung 5.2: Aufteilung der Verfahren mit Beispielen und Zuordnung der folgenden Abschnitte und Kapitel.

Abbildung 5.2 gibt dazu eine Übersicht mit einigen aus der Literatur bekannten Beispielen. Zunächst werden in den Kapiteln 6 bis 8 jeweils bekannte Verfahren betrachtet und evaluiert. Die Schwachpunkte werden durch Änderungen oder Erweiterungen behoben, was jeweils zu neuen Verfahren führt. Im Kapitel 9 werden die Konzepte der Störkompensation und der Modelanpassung schließlich in einem neuen Ansatz kombiniert, um einen Ausgleich aller relevanten Störungen zu schaffen. So werden die noch bestehenden Nachteile der einzelnen Methoden behoben und die Vorteile der verschiedenen Ansätze geeignet miteinander verschmolzen.

5.2 Ausreißerbehandlung

Beschränkung von Extremwerten

Bei abweichender Einsatzumgebung mit gestörten Eingangsdaten kommt es zwangsläufig zu **Ausreißern**, die nicht durch die Modellbildung in der Trainingsumgebung berücksichtigt sind. Obwohl diese Ausreißer vielleicht nur lokal beschränkt sind, können ihre Auswirkung beträchtlich sein.

Nehmen wir an, ein Merkmalsvektor $\mathbf{x}[k]$ innerhalb eines Lautes (Zustand z_s) sei sehr untypisch. Die Dichteschätzung $p(\mathbf{x}[k] \mid z_s, \lambda)$ des akustischen Modells λ für diesen Laut ist daher nicht sehr zuverlässig. Im Randbereich der typischerweise verwendeten Gaußmixturen dominiert dann meist eine Gaußdichte m. Bei Verwendung diagonaler Kovarianzmatrizen ergibt sich daher für die lokale Bewertung:

$$\begin{aligned} d(\mathbf{x}[k], z_s) &= -\log\left(p(\mathbf{x}[k] \mid z_s, \lambda)\right) \\ &\approx -\log\left(N(\mathbf{x}[k]; \boldsymbol{\mu}_{s,m}, \mathbf{D}_{s,m})\right) = -\log\left(\prod_i N(x_i[k]; \mu_{i;s,m}, \sigma^2_{i;s,m})\right) \\ &= \sum_i k_{i;s,m} + \sum_i \frac{(x_i[k] - \mu_{i;s,m})^2}{2\sigma^2_{i;s,m}} \end{aligned} \tag{5.1}$$

Hier ist $k_{i;s,m}$ ein Wert, der nicht vom Merkmalsvektor abhängt. Die Koeffizienten $x_i[k]$ gehen quadratisch in die Berechnung ein. Die akustische Bewertung der Zustandskette für das übergeordnete Wort beruht auf der Summe der lokalen Bewertungen. Weicht nur ein Koeffizient sehr stark ab, dominiert er durch den quadratischen Zusammenhang die gesamte Bewertung. In [Veth *et al* 1998, 1999] wird daher ein akustischer Rückfallwert (engl.: *backing-off*) vorgeschlagen, der die Auswirkung stark abweichender Muster begrenzen soll (siehe **Abbildung 5.3**).

Ähnliche Überlegungen lassen sich zum Beispiel auch für die Dauer von Phonemen oder Wörtern führen [Gu *et al.* 1991, Kim *et al.* 1994, Laurila 1997]. Gerade weil oft auf eine genauere Modellierung der Übergangswahrscheinlichkeiten verzichtet wird, kann eine Beschränkung der zeitlichen Ausdehnung dieser Einheiten manche Fehler verhindern. Ob jedoch durch den Ausschluss bestimmter Extremwerte eine Verbesserung erreicht werden kann oder die so verhinderten Fehler nur durch andere ersetzt werden, hängt von vielen Faktoren ab.

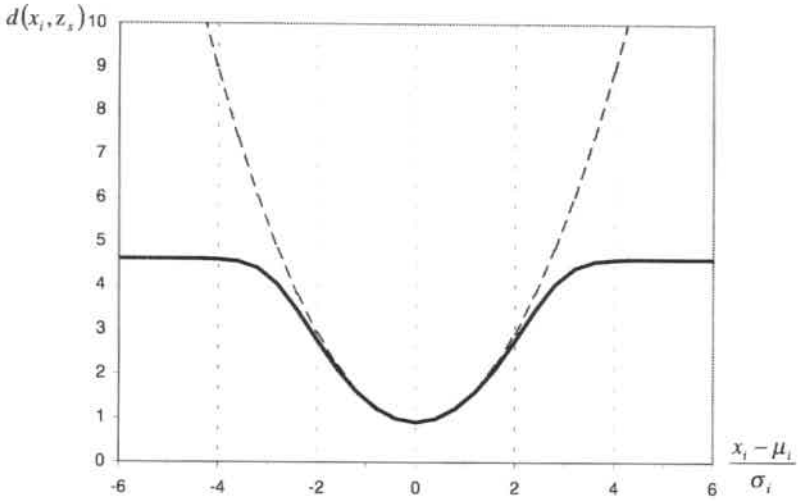


Abbildung 5.3: Beitrag einer lokalen Bewertung ohne (gestrichelt) und mit (durchgezogen) akustischem Rückfall.

Missing Feature Theory

Bei der *Missing Feature Theory*, zu der in jüngster Zeit einige Veröffentlichungen im Bereich Spracherkennung erschienen sind ([Cooke *et al.* 1997, Lippman und Carlson 1997] und einige mehr in den folgenden Jahren), werden ebenfalls Ausreißer behandelt. Einzelne Merkmalskoeffizienten $x_i[k]$ von bestimmten Analyserahmen k werden dabei explizit als fehlend bzw. nicht verlässlich gekennzeichnet. Um die daraus entstandenen Verfahren besser beschreiben zu können, werden die Koeffizienten eines teilweise gestörten Merkmalsvektors $\mathbf{x}[k]$ umsortiert und in einen Teil mit vorhandenen Merkmalen $\mathbf{x}_p[k]$ (engl.: *present*) und einen anderen mit fehlenden Merkmalen $\mathbf{x}_m[k]$ (engl.: *missing*) zusammengefasst.

$$\mathbf{x}[k] \rightarrow \begin{pmatrix} \mathbf{x}_p[k] \\ \mathbf{x}_m[k] \end{pmatrix} \tag{5.2}$$

Diese Sortierung erfolgt für jeden Analyserahmen unterschiedlich, je nachdem welche Merkmale gestört sind. Darauf basierend haben sich zwei weitere Vorgehen herauskristallisiert: Die fehlenden Merkmale

1. werden bei der Auswertung nicht oder nur unscharf berücksichtigt oder

2. werden rekonstruiert.

Einen Ansatzpunkt für das erste Vorgehen gibt die lokale Bewertung, wie sie in Gleichung (5.1) dargestellt wurde. Die Beiträge der einzelnen Koeffizienten konnten voneinander getrennt werden. Ist der Koeffizient unzuverlässig (*missing*), wird sein Beitrag bei allen

Auswertungen gestrichen und beeinflusst das Ergebnis auf diese Art überhaupt nicht. Dies entspricht einer Unschärfe dieses Koeffizienten über den gesamten Merkmalsbereich, weshalb man auch seinen Dichtebeitrag durch die entsprechende Randverteilung ersetzten kann. Das Ergebnis dieser Integration ist definitionsgemäß 1 und daher auch das Verschwinden des Beitrages zur lokalen Bewertung theoretisch fundiert:

$$d(x_i[k], z_s, m) = -\log\left(\int_{-\infty}^{\infty} \mathrm{N}(x_i[k]; \mu_{i;s,m}, \sigma^2_{i;s,m}) dx_i\right) \quad \text{für } x_i[k] \text{ aus } \mathbf{x}_{\mathrm{m}}[k] \qquad (5.3)$$
$$= -\log(1) = 0$$

Liegt etwas mehr Information über den gestörten Koeffizienten vor, kann er beispielsweise auf einen Bereich zwischen $-\infty \ldots x_i^{\max}$ eingegrenzt werden, so kann das Integral auch über diese Grenzen bestimmt werden. Bei [Renevey und Drygajlo 1999] wird der mögliche Bereich des unzuverlässigen Merkmals durch eine Gaußdichte beschrieben, die aus einer Geräuschschätzung hervorgeht. Der Beitrag ergibt sich dann über das Integral des Produktes dieser Gaußdichte und der Dichte für das Modell.

Die zweite Art vorzugehens ist eine Rekonstruktion der fehlenden Merkmale. Das einfache Nullsetzen oder Gleichsetzen mit dem Mittelwert des Merkmals wurde zwar auch in Betracht gezogen, erscheint jedoch etwas zu stark vereinfachend. [Raj *et al.* 1998] schlagen eine Interpolation der fehlenden Spektral-Merkmale aus den angrenzenden ungestörten Merkmalen vor. Wie jedoch die Analysen in Kapitel 8 zeigen, sind bei verrauschten Daten gerade kleine Spektralwerte gestört und ließen sich aus den jeweils energiestärkeren Nachbarkoeffizienten in der Praxis nur schlecht interpolieren. Eine etwas aufwändigere Methode benutzt einen linearen Schätzer (siehe Abschnitt 3.3.2), der mithilfe von Korrelationsmatrizen aus den ungestörten Merkmalen die fehlenden bestimmt:

$$\hat{\mathbf{x}}_{\mathrm{m}}[k] = \boldsymbol{\mu}_{\mathrm{m}} + \mathbf{C}_{\mathrm{m,p}} \cdot \mathbf{C}_{\mathrm{p,p}}^{-1} \cdot \left(\mathbf{x}_{\mathrm{p}}[k] - \boldsymbol{\mu}_{\mathrm{p}}\right) \qquad (5.4)$$

$$\text{mit } \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_{\mathrm{p}} \\ \boldsymbol{\mu}_{\mathrm{m}} \end{pmatrix} \text{ und } \mathbf{C} = \begin{pmatrix} \mathbf{C}_{\mathrm{p,p}} & \mathbf{C}_{\mathrm{p,m}} \\ \mathbf{C}_{\mathrm{m,p}} & \mathbf{C}_{\mathrm{m,m}} \end{pmatrix} \qquad (5.5)$$

Die Anwendung der *Missing Feature Theory* setzt die Kenntnis der gestörten Merkmale voraus. Dies wird meist mit einer Schwelle entschieden, die von der Stärke des geschätzten, additiven Rauschens abhängt. Dabei wird eine binäre Entscheidung getroffen, entweder gestört (also komplett wertlos) oder nicht gestört (also brauchbar zur Schätzung oder Auswertung), was eine sehr grobe Kategorisierung darstellt. Die Wahl des Merkmalsraumes für das akustische Modell ist auf spektrale Merkmale beschränkt. Da Cepstren auf allen Spektralkoeffizienten basieren, kann nicht entschieden werden welche Koeffizienten fehlen und welche nicht. Auch bei den Delta-Koeffizienten ergeben sich gewisse Probleme, denn auch sie können nicht berechnet werden, sobald einer der zugrundliegenden Koeffizienten fehlt.

Der Gedanke der Rekonstruktion wird dennoch auch in der vorliegenden Arbeit verfolgt, wenn auch aus der Sichtweise der Störkompensation. Während bei der *Missing Feature Theory* eine explizite Bestimmung der fehlenden Merkmale benötigt wird, geschieht dies in dem weiter unten vorgestellten Verfahren (MAM, siehe Kapitel 9) auf implizite Weise.

5.3 Robuste Verarbeitung

Auch wenn es gelingt die Erkennungsrate eines automatischen Spracherkenners auf das Niveau menschlicher Leistung oder sogar darüber hinaus zu bringen, gibt es dennoch Situationen bei denen Fehler unvermeidbar sind. Menschen versprechen sich zuweilen oder benutzen irreguläre grammatikalische Konstruktionen. Verschleifungen von Wörtern oder Häsitationen sind weitere Merkmale spontaner Sprache. Besonders Nebengeräusche des Sprechers oder der Umgebung führen bisweilen zu Einfügungen irrtümlich erkannter Wörter.

Effekte, die durch den spontanen Charakter der Äußerungen auftreten, werden bei dem verwendeten Basiserkenner zum Teil modelliert. Zum einen werden neben regulären Wörtern im Vokabular auch Geräusche (vom Menschen und der Umgebung) berücksichtigt [Ward 1989, Schultz 1994]. Ein lautes Atmen, eine Häsitation (wie „äh") oder Papierrascheln wurden in den GSST-Daten verschriftet und dann als eigene Modelle trainiert. Gerade aber kurze Umgebungsgeräuschen sind stark vom Umfeld abhängig (Beispiel Blinker im Auto) und können daher nicht allgemein erfaßt werden (zumal sich diese Geräusche auch mit der Sprache überlagern können).

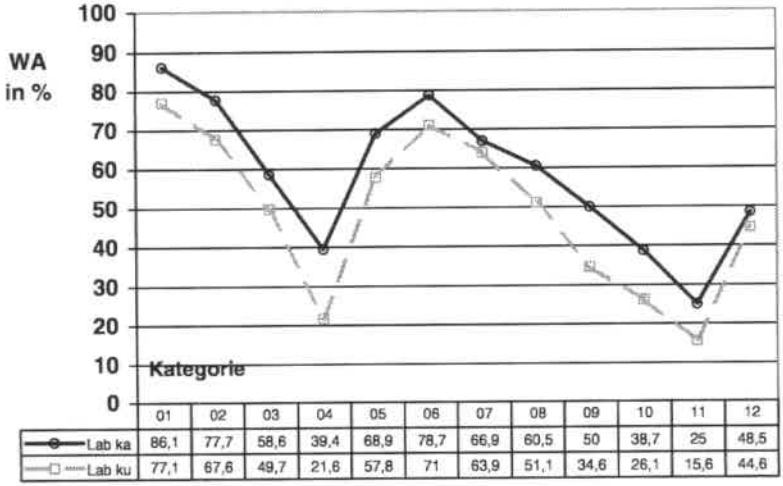
Weitere Maßnahmen wurden wegen der Verschleifung von Wörtern getroffen. Diese treten besonders häufig bei gängigen Formulierungen oder oft benutzten Funktionswörtern auf und hängen natürlich auch von regionalen Unterschieden ab. Eine spontansprachliche Anfrage klingt dann zum Beispiel so:

„Wie komm´ ich ´n zum Haub'bahnhof?"

Für Wörter wie „komme" und „denn" gibt es daher im Wörterbuch Aussprachevarianten.

Koartikulationseffekte (wie im Beispiel „Hauptbahnhof") werden durch die kontextabhängigen Phonemmodelle berücksichtigt. Hier stellt sich allerdings die Frage, ob eine Modellierung sehr vieler feiner Modelle (im vorliegenden Fall sind es 2.500) auch für gestörte Umgebungen sinnvoll ist. Womöglich hilft die Unterscheidung des Kontextes nur bei ungestörten Daten, während bei Störungen gröbere Klassen geeigneter sind. Wie oben bei der Ausreißerbehandlung würden so untypische Muster weniger ins Gewicht fallen, da sie durch das gröbere Modell eher noch erfasst werden. Diese Frage wurde experimentell geklärt.

Abbildung 5.4 zeigt die Ergebnisse des Basiserkenners, also mit geballten kontextabhängigen Phonemmodellen, im Vergleich mit einem kontextunabhängigen System, das lediglich knapp 200 akustische Modelle für die verschiedenen Phonemteile unterscheidet. Wie auch für Labordaten schneidet der kontextabhängige Erkenner für alle Kategorien besser ab. Damit ist der größere Aufwand, der in die akustische Modellierung gesteckt wurde, trotz erhöhtem Speicher- und Rechenzeitbedarf gerechtfertigt.



	01	02	03	04	05	06	07	08	09	10	11	12
Lab ka	86,1	77,7	58,6	39,4	68,9	78,7	66,9	60,5	50	38,7	25	48,5
Lab ku	77,1	67,6	49,7	21,6	57,8	71	63,9	51,1	34,6	26,1	15,6	44,6

Abbildung 5.4: Kontextabhängige (ka) und kontextunabhängige (ku) Phonemmodelle.

Bei starken Störeinflüssen häufen sich auch die Fehlerkennungen bei rein menschlicher Kommunikation. Die fehlende Information wird in solchen Fällen entweder durch Nachfragen in Erfahrung gebracht oder möglichst sinnvoll ergänzt. Bei einer automatisch erkannten Wortfolge hängt es von der jeweiligen Applikation ab, wie verfahren werden kann. Für eine Diktieranwendung ist jedes Wort wichtig und so müssen geeignete Korrekturmethoden zur Verfügung stehen [Suhm *et al.* 1996, Suhm 1998]. Bei spontansprachlichen Anfragen an ein Informationssystem sind manche Erkennungsfehler überhaupt nicht relevant. Die häufigen „den"/„dem"-Verwechsler sind bei den hier betrachteten Anfragen nicht für das Verständnis einer Anfrage entscheidend. Auch eingefügte Wörter sind nicht weiter tragisch, da sie von der nachfolgenden Sprachverarbeitungskomponenten mit den restlichen Wörtern nicht in Zusammenhang gebracht werden können und daher meist ignoriert werden.

Ein intelligentes sprachgesteuertes Informationssystem sollte in der Lage sein bei zweifelhaften Anfragen (sei es durch Falscherkennung oder unpräziser Eingabe) in einen Dialog einzutreten und den Sachverhalt zu klären. Zumindest muss eine Möglichkeit gegeben werden, eine ungewollte Reaktion auf eine falsche Eingabe mit einfachen Mitteln wieder rückgängig zu machen. Konfidenzmaße können dabei helfen sprachliche Eingaben als „fragwürdig" zu erkennen. Menschen verlassen sich bei der Interpretation des Gehörten oft auf den Kontext (wie Dialogzustand, zuvor erwähnte Themen, Weltwissen). Dieses Wissen ist für einen Spracherkenner meist nicht zugänglich, könnte aber helfen die korrekte Erkennung (oder Interpretation) des gesprochenen Wortes besser zu gewährleisten.

5.4 Robuste Merkmale

Die in der vorliegenden Arbeit verwendeten Merkmale sind bereits stark auf den Gebrauch für einen kontuierlichen Spracherkenner optimiert. Neben der kompakten Darstellung der sprachlich übermittelten Information besitzen sie bereits Eigenschaften, die sie gegenüber vielen Störungen robust machen. Obwohl auch immer wieder alternative Verfahren entwickelt werden (z.B. PLP [Hermansky 1990], RASTA [Hermansky und Morgan 1994]), haben sich die hier verwendeten Merkmale bewährt und sollen daher lediglich bezüglich ihrer Eigenschaften diskutiert werden. Beim RASTA-Filter (von engl.: *RelAtive SpecTrAl*) handelt es sich im Grunde um eine schritthaltende Mittelwertsubtraktion, wie sie als Störkompensation auch in [Wittmann *et al.* 1993] beschrieben wird. Diese Art der schritthaltenden Kanalschätzung lieferte jedoch in Vorversuchen zu dieser Arbeit schlechtere Ergebnisse als eine äußerungsbasierte Schätzung (siehe auch [Veth und Boves 1996]). In [Hermansky *et al.* 1993] wurde RASTA für additives Rauschen erweitert, womit es sich der weiter unten behandelten Spektralen Subtraktion annähert. Das Verfahren macht die Merkmale jedoch sehr stark vom zeitlichen (phonetischen) Kontext abhängig, was bei der hier verwendeten phonembasierten Spracherkennung von Nachteil ist.

5.4.1 Dynamische Merkmale und alternative Merkmalsräume

Die zeitlichen Ableitungen der Merkmalskoeffizienten (siehe Seite 22) haben sich wiederholt als hilfreich für die Spracherkennung erwiesen, weshalb dies hier nicht nochmals experimentell bestätigt werden muss. Diese dynamischen Merkmale sind unempfindlich gegenüber Kanaländerungen, wenn man sie wie hier von cepstralen Merkmalen berechnet. Alle zeitlich konstanten, im Cepstralbereich additiven Beiträge (wie der Kanal, siehe Abschnitt 8.2) werden durch die Differenzbildung entfernt. Während die dynamischen Merkmale vor allem zur Identifikation von Übergängen und kurzen, veränderbaren Lauten (wie z. B. den Plosiven) dienen, sind zur Unterscheidung von relativ stationären Lauten (z. B. Vokale) auch die statischen Merkmale wichtig.

Alternativ zu cepstralen Merkmalen werden teilweise auch logarithmierte, spektrale Merkmale verwendet, die durch eine Filterbank zusammengefasst werden. Wegen der Überlappung der Filter im Frequenzbereich und die immer noch relativ hohe Frequenzauflösung sind diese Log-Spektralkoeffizienten allerdings stark korreliert. Durch die cepstrale Analyse werden die Koeffizienten recht gut dekorreliert und dadurch kompakter darstellbar. Praktisch wird das Cepstrum aus einer dimensionsreduzierenden Matrixmultiplikation gewonnen. Die oben behandelten „*Missing Features*" werden dadurch nicht mehr identifizierbar. Die additive Störung, die sich vor allem bei energieschwachen Koeffizienten bemerkbar macht, wird auf alle Cepstralkoeffizienten verteilt. Falls man keine Maßnahmen gegen diese Störungen unternimmt, ist dieses Verhalten jedoch vorteilhaft.

Eine sehr geeignete Methode auch verschiedene Merkmale zu kombinieren und sie dabei durch Redundanzreduktion sehr kompakt darzustellen, bietet die LDA. Die resultierenden Merkmale sind dekorreliert und normiert und im Unterschied zu anderen Dimensionsreduktionsverfahren (wie die Hauptachsentransformation oder *principle component analysis*) wird bei der Auswahl der Merkmale die Unterscheidbarkeit der Klassen berücksichtigt.

5.4.2 Lineare Diskriminanzanalyse (LDA)

Grundlagen

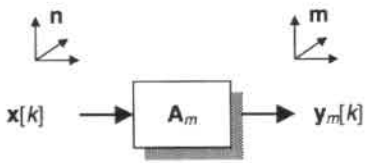


Abbildung 5.5: Lineare Abbildung zur Dimensionsreduktion mit $m < n$.

Mit der Linearen Diskriminanzanalyse wird eine Transformationsmatrix **A** bestimmt, mit der die n-dimensionalen Merkmalsvektoren **x** in die Merkmalsvektoren **y** abgebildet werden. Deren ersten m Koeffizienten $\mathbf{y}_m$ (wobei m beliebig gewählt werden kann) ergeben den maximalen Wert für $LD = \left|\mathbf{W}_m^{-1} \cdot \mathbf{T}_m\right|$ (siehe Abschnitt 3.3.3), der mit einer linearen Transformation in einen m-dimensionalen Raum erreichbar ist.

Die Varianzen der Koeffizienten $\mathbf{y}_m$ sind im Vergleich zu den Varianzen innerhalb der Klassen, zu denen die Muster zugeordnet sind, maximiert. Die Koeffizienten von **y** sind gewissermaßen nach ihrer Diskriminanz sortiert, so dass man durch Auswahl einer $m \times n$-Teilmatrix $\mathbf{A}_m$ eine gewünschte Anzahl von geeigneten Merkmalen erhält (**Abbildung 5.5**). Die Lösung für die Matrix **A**, deren Herleitung man z.B. in [Fukanaga 1972] findet, ist:

$$\mathbf{A} = (\phi_1 \phi_2 \ldots \phi_n)^T, \quad \text{mit} \quad \lambda_1 > \lambda_2 > \ldots > \lambda_n \tag{5.6}$$

Dabei sind ϕ_i die normierten Eigenvektoren von $\mathbf{W}^{-1}\mathbf{T}$, die nach der Größe ihrer Eigenwerte λ_i geordnet wurden.

Eine Anschauung dieser Transformation erschließt sich über die *simultane Diagonalisierung* (ebenfalls in [Fukunaga 1972] zu finden), die nun kurz dargestellt wird. Dabei wird die Transformationsmatrix **A** aus einzelnen Teilabbildungen zusammengesetzt (**Abbildung 5.6**). Da eine Skalierung der Koeffizienten nichts am LD-Wert ändert, werden die Muster **x** zunächst so transformiert und skaliert, dass die mittlere Klassen-Kovarianzmatrix **W** zur Einheitsmatrix wird. Stellt man die Matrix **W** wie in Abschnitt 3.3.3 für den zweidimensionalen Fall als Ellipse dar, werden ihre Hauptachsen dabei parallel zu den Koordinatenachsen ausgerichtet und auf 1 normiert. Dazu werden die Muster mit einer Matrix bestehend aus den Eigenvektoren von **W** multipliziert (Achsenparallelisierung) und dann durch eine weitere Matrix mit den reziproken Eigenwerten in der Diagonale, in Achsenrichtung normalisiert. Durch diese erste Stufe der Transformation ist die neue mittlere Klassen-Kovarianzmatrix eine Einheitsmatrix und ändert ihre Gestalt durch die abschließende Achsenparallelisierung der Kovarianzmatrix **T** nicht mehr. Die größte Diskriminanz besitzen nun die Koeffizienten von **y**, die den größten Eigenwert der resultierenden Kovarianzmatrix **T** vorweisen.

Einsatz in der Spracherkennung

Für die Spracherkennung hat der Einsatz der Linearen Diskriminanzanalyse eine Reihe von Vorteilen:

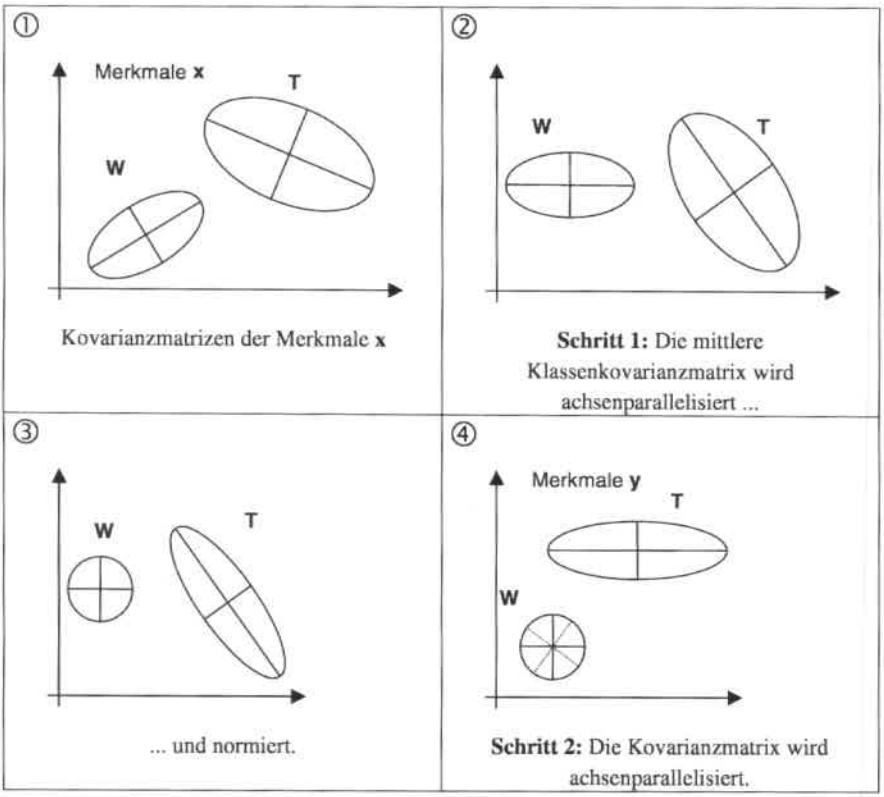


Abbildung 5.6: Simultane Diagonalisierung.

- Dekorrelation der Merkmale
- Normierung der Merkmale
- Redundanzreduktion der Merkmale durch freie Wahl ihrer Anzahl

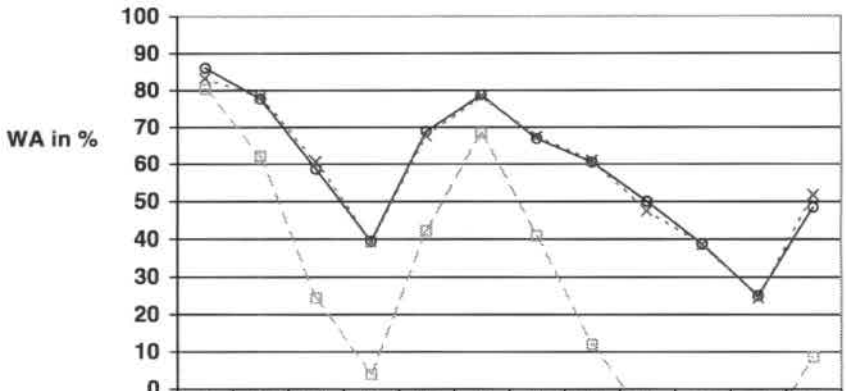
So können auch unterschiedliche Merkmale verwendet oder neue, unter Belassung der Parameteranzahl des akustischen Modells, hinzugefügt werden. Die Auswahl geeigneter Merkmale geschieht gewissermaßen automatisch, denn Rauschen und redundante Information wird bei der Reduktion herausfallen. Bestimmte, über einen linearen Zusammenhang abgeleitete Merkmale können implizit über die LDA realisiert werden. Die Cepstren gehen z.B. aus einer linearen Transformation aus den Log-Spetralmerkmalen hervor und die dynamischen Merkmale aus benachbarten Merkmalsvektoren. Verwendet man also diese Merkmale als Ausgangsmerkmale **x**, können die genannten Merkmale theoretisch aus der LDA-Transformation hervorgehen.

Die Frage nach einer geeigneten Wahl der Klassen für die LDA konnte in [Haeb-Umbach und Ney 1992, Aubert *et al.* 1993, Maier 1994] geklärt werden. Auch wenn man prinzipiell nur an der Unterscheidung der Phoneme interessiert ist, erwiesen sich kontextabhängige

Phonemteile als vorteilhaftere Klassen. Bei den hier verwendeten Erkennern wurden die geballten Einheiten (2.500 Stück) verwendet.

Wie weiter oben angedeutet, verhindert die Transformation in den Cepstralbereich bzw. die LDA-Transformation, eine Identifizierung einzelner durch Rauschen besonders gestörter Koeffizienten (*Missing Features*). Es stellt sich daher die Frage, wie stark sich diese Transformationen auf die Erkennungsleistung auswirken und ob man eventuell auch auf sie verzichten kann. **Abbildung 5.7** zeigt die Ergebnisse von Systemen mit verschiedenen Merkmalen für das akustische Modell. Beim Vergleich der beiden oberen Systeme mit *Mel*-Cepstralkoeffizienten (MFCC[14]) bzw. *Mel*-Spektralkoeffizienten (MFSC[15]) ergeben sich keine großen Unterschiede. Dies ist wohl auf die verwendete LDA-Transformation zurückzuführen, die die hier bestehenden Unterschiede größtenteils ausgleichen kann. Verwendet man die reinen MFSC samt ihrer dynamischen Merkmale ohne LDA, fallen die Erkennungsraten deutlich geringer aus. Die Einbußen treten besonders stark bei den sowieso schon schlechter abschneidenden Kategorien auf.

Die LDA trägt also auch zur Robustheit gegenüber den hier auftretenden Störungen bei. Auch wenn das System ohne LDA die Möglichkeit bietet, unzuverlässige Koeffizienten auszumachen, beträgt die Einbuße gegenüber dem Basiserkenner auch bei der kaum gestörten Kategorie 01 schon knapp 6 % absolut und wesentlich mehr bei anderen Kategorien. Damit erweist sich die LDA als wichtiger und nicht verzichtbarer Bestandteil zur robusten Erkennung.



Besonderheit	01 aus	02 Blinker	03 Fenster	04 Lstark	05 Lmittel	06 0km/h	07 25km/h	08 50km/h	09 75km/h	10 100km/	11 125km/	12 beschl.
—o— MFCC, LDA	86,1	77,7	58,6	39,4	68,9	78,7	66,9	60,5	50	38,7	25	48,5
- - ✕ - - MFSC, LDA	83,3	78,1	60,6	39,3	67,6	78,4	67,5	61	47,5	38,7	24,4	51,8
— ⊟ — MFSC, noLDA	80,3	62,1	24,4	4	42,3	68,5	41	12	-8,5	-13,2	-14,5	8,6

Abbildung 5.7: Vergleich zwischen cepstralen (MFCC) und log-spektralen (MFSC) Merkmalen mit LDA und log-spektralen Merkmalen ohne LDA.

[14] von engl.: *mel frequency cepstral coefficients*
[15] von engl.: *mel frequency spectral coefficients*

6 Modellanpassung

Alles Vernünftige hat seinen Ausgangspunkt und sein Ziel.
Die Bewegung ist die Äußerung des Lebens.
- Belenki, russischer Autor -

6.1 Überblick

Hier soll der Begriff „Modellanpassung" (auch Modelladaption) zwar auf das akustische Modell beschränkt bleiben, dennoch in einem erweiterten Kontext zur sonstigen Verwendung in der Literatur betrachtet werden. Ausgangspunkt ist ein Spracherkennungssystem, das mit Beispielen aus einer Umgebung A trainiert wurde, und nun in einer Umgebung B verwendet werden soll (siehe **Abbildung 6.1**).

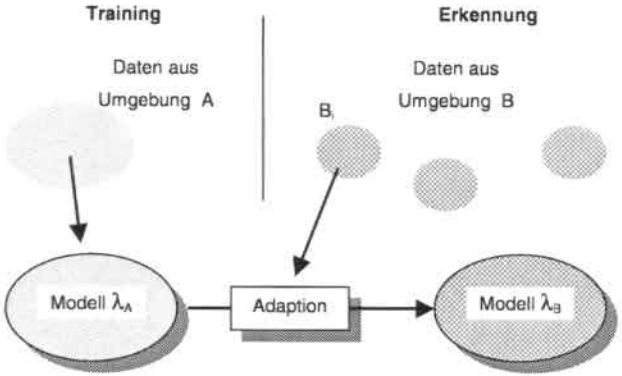


Abbildung 6.1: Anpassung an eine neue Umgebung

Mit Umgebung sei, wie auch schon im einführenden Kapitel behandelt, vor allem das akustische Umfeld sowie wechselnde Sprecher gemeint. In einigen Fällen wird sich vielleicht sogar die Art der Anwendung (*Szenario*) völlig ändern. Dies könnte zum Beispiel bewirken, dass der bevorzugte Sprechstil sich von Diktiereingaben in spontansprachliche Anfragen ändert oder von kontinuierlicher Eingabe in isolierte Kommandowörter. Einige Parameter der vom System verwendeten Modelle stimmen nun nicht mehr mit den neuen Gegebenheiten überein, was zu Leistungseinbußen führt. Dies trifft im vorliegenden Fall besonders für die Dichteschätzungen der phonetischen Klassen zu. Änderungen des (linguistischen)

Sprachmodells, des Wörterbuchs oder des Entscheidungsbaums zur Einteilung in phonetische Klassen werden hier nicht betrachtet.

Einmalige Anpassung für eine neue Anwendung

Liegen ausreichend verschriftete Trainingsdaten der neuen Umgebung vor und bleiben die Bedingungen dort relativ konstant, lässt sich die Anpassung am besten dadurch vornehmen, indem die Dichteparameter des akustischen Modells völlig neu geschätzt werden (**Neutraining**). Auch für unterschiedliche Bedingungen lassen sich so meist relativ gute Ergebnisse erzielen. Um beispielsweise einen sprecherunabhängigen Erkenner zu erhalten, werden als Trainingsdaten Aufnahmen möglichst vieler unterschiedlicher Sprecher verwendet (wie auch hier geschehen). Die Varianz innerhalb der phonetischen Klassen wird dadurch aber größer. So schneiden sprecherunabhängige Erkenner trotz einer größeren Menge an Trainingsdaten meist schlechter ab als sprecherabhängige. Oft lohnt es die Trainingsmenge zu teilen und jeweils ein gesondertes akustisches Modell für Frauen und Männer zu trainieren. Ähnlich könnte man auch für die unterschiedlichen Rauschkategorien usw. argumentieren. Nur verringert sich dadurch die Trainingsmenge immer weiter und es muss auch möglich sein, zwischen diesen Konditionen während der Erkennung zu unterscheiden.

Um für bestimmte Arten von Störungen Trainingsmaterial zu beschaffen, kann man auf die meist in großer Menge zur Verfügung stehenden Sprachdaten aus der ungestörten Laborumgebung zurückgreifen. So werden teilweise um Erkenner für Telefonanwendungen zu erstellen, Labordaten über Telefonleitungen oder entsprechende Simulatoren übertragen und wieder neu aufgenommen . Liegen wie für diese Arbeit Aufnahmen des Hintergrundgeräusches vor, kann man damit die Labordaten nachträglich durch additive Überlagerung künstlich verrauschen und damit den Erkenner ebenfalls neu trainieren (**Neutraining mit simulierten Daten**). Dies hat gegenüber den „echten" Daten noch den Vorteil, dass man die Zuordnung der Muster zu den phonetischen Klassen über die ungestörten Daten und das entsprechende Modell λ_A vornehmen kann (**Abbildung 6.1**).

Um den Trainingsprozess mit verrauschten Daten zu beschleunigen, kann man statt der Trainingsdaten ein bereits existierendes akustisches Modell λ_A für ungestörte Sprache verwenden. Dies ist ja nichts anderes als eine sehr kompakte Darstellung der zur Erkennung wichtigen akustischen Eigenschaften der Trainingsmenge. Dieses Modell kann man in bestimmten Fällen direkt durch Kombination mit einem akustischen Modell für Rauschen erstellen. In einem alternativen Vorgehen generiert man aus diesen beiden Modellen zufällig Sprachmuster, die dann kombiniert werden und zur Neuschätzung dienen. In beiden Fällen spricht man von (paralleler) **Modellkombination** (engl.: *parallel model combination*, PMC).

Die aufgeführten Möglichkeiten zur einmaligen Anpassung der akustischen Modelle werden in den folgenden beiden Abschnitten behandelt.

Abschnitt 6.2: Modellanpassung durch Neutraining

- Neutraining mit „echten" Daten
- Neutraining mit simulierten Daten

Abschnitt 6.3: Modellkombination

- Neubestimmung mit generierten Daten (Datengetriebene PMC)
- Neubestimmung durch Kombination (PMC)

Fortlaufende Anpassung

Für den hier betrachteten Fall der Autoumgebung besteht die neue Umgebung B aus vielen unterschiedlichen Situationen B_1 bis B_N. Außerdem soll der Erkenner nicht nur für diese Fälle, sondern auch für andere Umgebungen (Zug, Flugzeug, Fußgängerzone) gute Erkennungsraten liefern. Die Einschränkung, die daher für die hier gemachten Experimente gilt, erlaubt nur die Verwendung der aktuellen Aufnahme, da sich Sprecher und Geräuschsituation bei der nächsten Anfrage ändern können. Um bei der obigen Einteilung zu bleiben, liegen nun zwar „echte" Daten aus der neuen Umgebung vor, sie sind aber nicht verschriftet und ihre Menge ist äußerst gering. Deshalb können die Modelle nur zum Teil oder mit einer sehr eingeschränkten Transformation angepasst werden. Die dazu in Abschnitt 6.4 (Anpassung während der Erkennungsphase) aufgeführten Verfahren, speziell die MLLR, wurden vor allem zur Anpassung an den Sprecher entwickelt und werden meist mit mehr Adaptionsdaten als für die hier vorgestellten Experimente verwendet. Deshalb wird in diesem Abschnitt ein auf der MLLR aufbauender Ansatz entwickelt, der diesen Nachteil behebt (OSTD).

Abschnitt 6.4: Anpassung während der Erkennungsphase

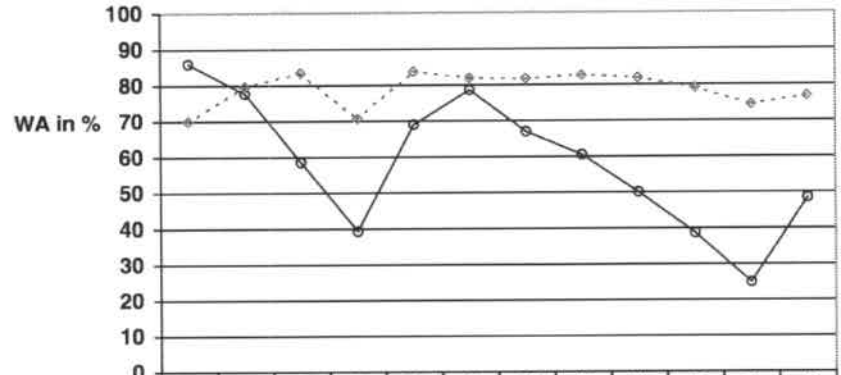
- Fortlaufende Anpassung mittels „echter" Daten durch lineare Regression (MLLR)
- Fortlaufende Anpassung mittels simulierter Daten durch lineare Regression (OSTD)

6.2 Modellanpassung durch Neutraining

6.2.1 Training mit Sprachdaten aus dem Auto

„Autospracherkenner" (Auto)

Abbildung 6.2 zeigt die Wortakkuratheit des Basiserkenners (Labor) aus Abschnitt 4.2, der nur mit Labordaten trainiert wurde, im Vergleich mit einem sonst gleichen System, für das Sprachaufnahmen aus der Autoumgebung verwendet wurden (Trainingsdaten aus Abschnitt 4.1.2). Der „Autospracherkenner" (Auto) liefert für die meisten Kategorien gute Ergebnisse. Für hohe Geschwindigkeiten (10 und 11) und Beschleunigen (12) ergibt sich ein leichtes Absinken, für starkes Lüftergeräusch (04) sind die Einbußen deutlicher. Etwas überraschend ist das schlechte Abschneiden bei Kategorie 01, bei der praktisch keine Hintergrundgeräusche vorhanden sind (Laborbedingung). Dies ist dadurch zu erklären, dass in den verwendeten Trainingsdaten kaum Beispiele mit dieser guten Signalqualität vorkommen, wodurch kein passendes akustisches Modell für diese Bedingung vorliegt.



	01	02	03	04	05	06	07	08	09	10	11	12
Besonderheit	aus	Blinker	Fenster	L.stark	L.mittel	0km/h	25km/h	50km/h	75km/h	100km/	125km/	beschl.
Labor	86,1	77,7	58,6	39,4	68,9	78,7	66,9	60,5	50	38,7	25	48,5
Auto	69,9	79,7	83,5	70,6	83,9	82	81,7	82,7	82	79,1	74,4	76,8

Abbildung 6.2: Ergebnisse eines Erkenners der auf Laboraufnahmen trainiert wurde, im Vergleich mit einem auf Autosprachdaten trainierten Erkenner.

Für bestimmte Einzelfälle mag dieses Vorgehen zweckmäßig sein, es ergeben sich jedoch die folgenden Nachteile:

1. Eine umfangreiche Datensammlung (wie sie für das Neutraining nötig ist) ist langwierig und finanziell **aufwändig**. Alle Einsatzbedingungen und Sprechertypen müssen ausreichend abgedeckt sein. Je nach Art der Daten müssen diese manuell transliteriert werden.
2. Ein neues akustisches Modell muss trainiert werden, was ebenfalls einigen Aufwand darstellt.
3. Wenn die Umgebungsbedingungen erneut **abweichen** (B_{N+1}, z.B. neue Anwendung für Flugzeuge oder auch nur ein anderer Fahrzeugtyp), passt das Modell eventuell wieder nicht.
4. Ein Modell, das mit Daten aus verschiedenen Umgebungen B_1 bis B_N trainiert wurde, ist meist nicht optimal für den **Spezialfall** B_i.

Um zunächst Nachteil 1 zu reduzieren, wird im nächsten Abschnitt 6.2.2 untersucht wie mit einfacher zu beschaffenden reinen Geräuschaufnahmen Trainingsdaten simuliert werden können. Da dies keine Lösung der anderen Punkte mit sich bringt, werden später Verfahren vorgestellt, mit denen die Modelle während der Erkennung auf die aktuelle Situation eingestellt werden. Für diese und die Experimente der noch folgenden Kapitel werden die durch das Neutraining erzielten Ergebnisse jedoch als Referenz dienen.

6.2.2 Training mit simulierten Daten

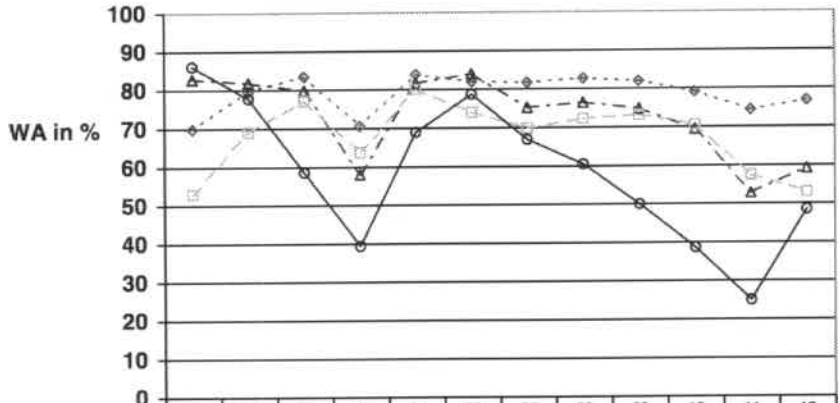
Neben den in Abschnitt 4.1.2 beschriebenen Sprachdaten aus der Autoumgebung wurden bei dieser Sammlung auch diverse reine Geräuschaufnahmen gemacht. Im Vergleich mit den Sprachdaten sind diese sehr viel einfacher auch in großen Mengen zu beschaffen. Allerdings sei angemerkt, dass man kaum sicherstellen kann, alle Geräuschsituationen einer neuen Anwendungsumgebung ausreichend erfasst zu haben (Nachteil 3 im vorigen Abschnitt). Im Vergleich zum Training mithilfe realer Daten wird hier auch nur ein bestimmter Störaspekt berücksichtigt, nämlich additives Rauschen.

Erkenner Sim1

In einem ersten Versuch (Sim1) wurden zufällig Segmente aus 6 langen Geräuschaufnahmen (200 Sekunden, entspricht 20.000 Rauschmustern) von verschiedenen Fahrzeugen und Fahrsituationen den unter Laborbedingungen aufgenommenen GSST-Trainingsdaten zugemischt. Die Ergebnisse in **Abbildung 6.3** zeigen, dass besonders Kategorien mit hohen Rauschanteilen dadurch profitiert haben. Andererseits schneidet zum Beispiel Kategorie 01 wie auch schon beim Training mit den Autosprachdaten, sehr schlecht ab.

Erkenner Sim2

Daraufhin wurde ein zweites System (Sim2) trainiert, bei dem auch die Gewichtung der zugemischten Rauschsignale zufällig nach einer Gleichverteilung (Bereich zwischen 0 und 1) gewählt wurde. Damit konnte insbesondere die rauschfreie Kategorie 01 sowie die Kategorien 06 bis 08 (kleine Geschwindigkeiten) deutlich verbessert werden. Hier bestätigt sich die Tendenz, dass allgemeine Robustheit oft auf Kosten der Performanz für den Spezialfall geht (Nachteil 4), da die Ergebnisse für Kategorie 04 und 11 nun etwas schlechter ausfallen als bei „Sim1".



Besonderheit	01 aus	02 Blinker	03 Fenster	04 Lstark	05 Lmittel	06 0km/h	07 25km/h	08 50km/h	09 75km/h	10 100km/	11 125km/	12 beschl.
Auto	69,9	79,7	83,5	70,6	83,9	82	81,7	82,7	82	79,1	74,4	76,8
Sim2	82,8	81,8	79,8	58	81,6	84	75,2	76,4	74,8	69,4	52,8	59,2
Sim1	53	69,1	77,1	63,8	80	74,1	69,8	72,3	73,1	70,6	57,5	53,1
Labor	86,1	77,7	58,6	39,4	68,9	78,7	66,9	60,5	50	38,7	25	48,5

Abbildung 6.3: Ergebnisse mit künstlich verrauschten Trainingsdaten (Sim1, Sim2) im Vergleich mit realen und unverrauschten Daten (Auto und Labor).

6.2.3 Referenzsysteme

Mit den in **Abbildung 6.3** aufgezeigten Ergebnissen für die Systeme „Labor", „Sim2" und „Auto" können nun weitere Verfahren auf ihre Effektivität bezüglich einer robusten Erkennung in dieser Art von Umgebung beurteilt werden. Das Basissystem „**Labor**" stellt einen Erkenner dar, der in einer speziellen rauscharmen Umgebung trainiert wurde und dort auch die besten Ergebnisse liefert, aber an neuen Situationen scheitert. „Auto" und „Sim2" sind zwei komplett neu für die Autoumgebung trainierte Systeme. „**Auto**" liefert im Durchschnitt die deutlich besseren Ergebnisse, setzt aber eine aufwändige Datensammlung voraus. Die Datenbeschaffung für „**Sim2**" ist zwar wesentlich einfacher, dafür kann dieses System aber nur mit durchschnittlichen Ergebnissen aufwarten. Bei allen noch folgend vorgestellten Verfahren wurde versucht, in der Trainingsphase des Erkenners möglichst keine Daten aus der Autoumgebung zu verwenden, um auch für den direkten Einsatz in anderen ähnlich gearteten Umgebungen (Flugzeug, Zug, Fabrikhalle, Straße usw.) brauchbar zu sein. Dennoch werden die Ergebnisse stets mit den drei hier aufgeführten **Referenzsystemen** verglichen, um ihre Leistung besser einordnen zu können.

6.3 Modellkombination

Eine Möglichkeit bereits vorhandene Modelle ungestörter Sprache zu benutzen und wie beim Neutraining die Parameter für andere Umgebungen neu zu bestimmen, beschreitet die *Parallel Model Combination* (PMC) [Gales und Young 1992-1994, Gales 1995]. Im Gegensatz zur *Model Decomposition* [Varga und Moore 1990] ist das Ergebnis dieses Prozesses ein herkömmliches akustisches Modell, dass dann ohne weitere Modifikationen an den Suchalgorithmen des Erkenners verwendet werden kann.

Grundprinzip der Modellkombination

Ziel der PMC ist es, aus einem Modell für ungestörte Sprache mithilfe eines akustischen Modells für eine additive Störung ein kombiniertes Modell herzuleiten. Dies kann dann zur Erkennung von entsprechend verrauschter Sprache dienen.

Theoretisch kann das Modell der Störung aus einem HMM mit mehreren Zuständen bestehen [Martin *et al.* 1993], wodurch sich aber auch die Topologie des kombinierten Modells gegenüber dem ursprünglichen Sprach-HMM ändert (siehe **Abbildung 6.4**). In der Praxis wird daher meist nur mit einem Geräuschzustand gearbeitet, womit die ursprüngliche Topologie erhalten bleibt. Da man nicht wirklich an der Unterscheidung eventuell verschiedener Geräuschzustände interessiert ist, führt dies nur zum Verlust der Übergangsmodellierung zwischen den verschiedenen Geräuschen. Die Wahrscheinlichkeitsdichten zur Berechnung der Emissionswahrscheinlichkeiten können hingegen gemeinsam modelliert werden.

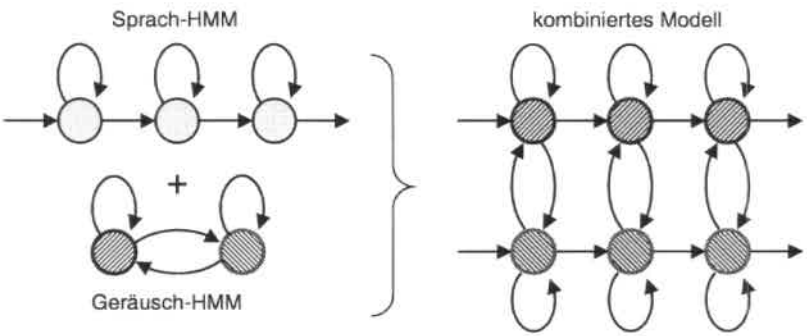


Abbildung 6.4: Änderung der Topologie bei der Modellkombination

Die Bestimmung der Dichten für das kombinierte Modell gestaltet sich mitunter recht aufwändig. Im einfachsten Fall wird das zur Sprache unkorrelierte Rauschen $\mathbf{n}$ als normalverteilt angenommen und das Rauschmodell λ_n besteht daher nur aus einer Gaußdichte, wird also durch einen Mittelwertvektor μ_n und die Kovarianzmatrix $\mathbf{C}_n$ repräsentiert. Werden durch dieses Modell Merkmale $\mathbf{n}$ beschrieben, die sich den Merkmalen $\mathbf{s}$ des Sprachmodells λ_s additiv überlagern, gilt also

$$\mathbf{x} = \mathbf{s} + \mathbf{n} \tag{6.1}$$

und ist **s** ebenfalls normalverteilt, so leitet sich daraus wiederum eine Gaußdichte für das neue Merkmal **x** mit den folgenden Eigenschaft ab:

$$\mu_x = \mu_s + \mu_n \tag{6.2}$$

$$\mathbf{C}_x = \mathbf{C}_s + \mathbf{C}_n \tag{6.3}$$

Dieser Zusammenhang gilt auch für die einzelnen Gaußdichten einer eventuellen Mischverteilung für λ_s. Für kompliziertere Geräuschmodelle geht man meist zu einem **datengetriebenen** Ansatz (**DPMC**, engl.: *data driven* PMC) über [Gales und Young 1995]. D.h. man generiert aus den zu kombinierenden Modellen zufällig Merkmalsvektoren, kombiniert diese und schätzt aus den so gewonnenen Trainingsbeispielen ein neues Modell.

Transformation der Modelle

Die additive Überlagerung trifft allerdings für die normalerweise zur Sprachmodellierung benutzten Merkmalsräume nicht zu. Spektrale Merkmale, für die dies näherungsweise gilt, werden meist logarithmiert und anschließend linear transformiert (Cepstrum und LDA, vergleiche Abschnitt 2.2 Merkmalsextraktion). Daher werden Modelle, die auf solchen Merkmalen aufbauen, in den spektralen Raum transformiert, dort kombiniert und anschließend wieder zurück transformiert. Es werden also Dichtefunktionen $p_y(\mathbf{y})$ benötigt, die aus einer Transformation der Merkmale $\mathbf{y} = f(\mathbf{x})$ hervorgehen. Der Zusammenhang mit der Dichte $p_x(\mathbf{x})$ kann veranschaulicht werden, wenn man sich klar macht, dass der lokale Dichteanteil $p_x(\mathbf{x})dx$ erhalten bleibt. Für den eindimensionalen Fall ergibt sich mit der stetigen und streng monotonen Funktion $f(x)$ und ihrer Umkehrfunktion $f^{-1}(y)$

$$\left|p_y(y)\cdot dy\right| = \left|p_x(x)\cdot dx\right| \quad \Rightarrow \quad p_y(y) = p_x(x)\cdot\left|\frac{dx}{dy}\right| = p_x(f^{-1}(y))\cdot\left|f^{-1}(y)'\right|. \tag{6.4}$$

Im mehrdimensionalen Fall wird bei der Umrechnung der Dichtefunktion die einfache Ableitung $f^{-1}(y)'$ durch die Jacobische Determinante ersetzt. Diese ist definiert als

$$\mathbf{J} = \frac{\partial(x_1,\ldots,x_n)}{\partial(y_1,\ldots,y_n)} = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \cdots & \frac{\partial x_n}{\partial y_1} \\ \vdots & & \vdots \\ \frac{\partial x_1}{\partial y_n} & \cdots & \frac{\partial x_n}{\partial y_n} \end{vmatrix} \tag{6.5}$$

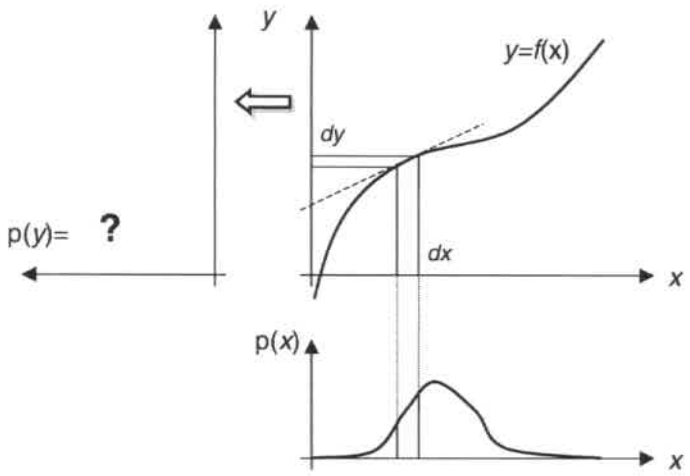


Abbildung 6.5: Transformation einer Dichtefunktion.

Beispiele:

1. $y = a \cdot x + b \quad \Rightarrow \quad p_y(y) = p_x\left(\frac{y-b}{a}\right) \cdot \frac{1}{|a|}$ (6.6)

2. $y = \ln(x) \quad \Rightarrow \quad p_y(y) = p_x(e^y) \cdot e^y$ (6.7)

Ist die ursprüngliche Dichtefunktion $p_x(x)$ normalverteilt, so ergibt sich nur für Beispiel 1 wieder eine Normalverteilung:

$$p_y(y) = \frac{1}{|a|} \cdot N\left(x;\mu_x,\sigma^2{}_x\right) = \frac{1}{|a|} \cdot N\left(\frac{y-b}{a};\mu_x,\sigma^2{}_x\right) = N\left(y;a\cdot\mu_x + b, a^2\cdot\sigma^2{}_x\right) \tag{6.8}$$

Für Beispiel 2 gilt dies offensichtlich nicht, man kann allerdings eine normalverteilte Näherung benutzen. Dazu werden der Mittelwert und die Varianz der exakten Darstellung bestimmt und man erhält:

$$p_y(y) \approx \hat{p}_y(y) = N\left(y;\mu_y,\sigma^2{}_y\right) \tag{6.9}$$

$$\mu_y = \ln\left(\frac{\mu_x{}^2}{\sqrt{\mu_x{}^2 + \sigma^2{}_x}}\right) \tag{6.10}$$

$$\sigma^2{}_y = \ln\left(\frac{\sigma^2{}_x}{\mu_x{}^2} + 1\right) \tag{6.11}$$

Eine Herleitung dieser Zusammenhänge für mehrdimensionale Gaußdichten findet man in [Gales 1995, Anhang C].

In **Tabelle 6.1** sind die einzelnen Schritte einer Modellkombination für cepstrale Merkmale zusammengefasst. In allen Merkmalsräumen werden Gauß'sche Mischverteilungen verwendet. Die Umrechnung der Parameter $\boldsymbol{\mu}$ und $\mathbf{C}$ wird jeweils für einzelne Gaußdichten gegeben. μ_i sind die Koeffizienten des Mittelwertvektors $\boldsymbol{\mu}$ und σ_{ij} die Matrixelemente der Kovarianzmatrix $\mathbf{C}$. Bis zur Kombination der beiden Modelle steht der Merkmalsvektor $\mathbf{x}$ stellvertretend für die unverrauschten Sprachmuster $\mathbf{s}$ wie auch für die reinen Geräuschmuster $\mathbf{n}$. Die hochgestellten Buchstaben bezeichnen den cepstralen (c), den logarithmischen (l) und den linearen (s) Spektralbereich.

Tabelle 6.1: PMC, Transformation und Kombination für Modelle mit cepstralen Merkmalen.

Ausgangspunkt Cepstrale Merkmale: $\mathbf{x}^c$ Modelle: λ_n^c; λ_s^c
Übergang in den logarithmierten Spektralbereich: $\mathbf{x}^l = \mathbf{A}_{cos}^{-1} \cdot \mathbf{x}^c$ $\boldsymbol{\mu}^l = \mathbf{A}_{cos}^{-1} \cdot \boldsymbol{\mu}^c \qquad \mathbf{C}^l = \mathbf{A}_{cos}^{-1} \cdot \mathbf{C}^c \cdot \left(\mathbf{A}_{cos}^{-1}\right)^T$ (6.12) Modelle: $\lambda_n^c \rightarrow \lambda_n^l$; $\lambda_s^c \rightarrow \lambda_s^l$
Übergang in den (linearen) Spektralbereich: $x_i^s = e^{x_i^l}$ *Näherung mit Normalverteilung:* $\mu_i^s = e^{\mu_i^l + \frac{1}{2}\sigma_{ii}^l} \qquad \sigma_{ij}^s = \mu_i^s \cdot \mu_j^s \cdot \left(e^{\sigma_{ij}^l} - 1\right)$ (6.13) Modelle: $\lambda_n^l \rightarrow \lambda_n^s$; $\lambda_s^l \rightarrow \lambda_s^s$

Kombination der Modelle: $\mathbf{x}^{\mathrm{s}} = \mathbf{s}^{\mathrm{s}} + \mathbf{n}^{\mathrm{s}}$

Bei nur einer Normalverteilung für das Rauschmodell:

$$\mu_x^{\mathrm{s}} = \mu_s^{\mathrm{s}} + \mu_n^{\mathrm{s}} \qquad \mathbf{C}_x^{\mathrm{s}} = \mathbf{C}_s^{\mathrm{s}} + \mathbf{C}_n^{\mathrm{s}} \tag{6.14}$$

Alternativ: **Datengetriebener Ansatz.**

Künstliche Beispiele werden aus den beiden Modellen generiert, addiert und dienen zur Schätzung des kombinierten Modells.

Modelle: $\lambda_s^{\mathrm{s}}, \lambda_n^{\mathrm{s}} \rightarrow \lambda_x^{\mathrm{s}}$

Übergang in den logarithmierten Spektralbereich: $x_i^{\mathrm{l}} = \ln(x_i^{\mathrm{s}})$

Näherung mit Normalverteilung:

$$\mu_i^{\mathrm{l}} = \ln \mu_i^{\mathrm{s}} - \frac{1}{2} \ln\left(\frac{\sigma_{ii}^{\mathrm{s}}}{(\mu_i^{\mathrm{s}})^2} + 1 \right) = \ln\left(\frac{(\mu_i^{\mathrm{s}})^2}{\sqrt{(\mu_i^{\mathrm{s}})^2 + \sigma_{ii}^{\mathrm{s}}}} \right)$$

$$\sigma_{ij}^{\mathrm{l}} = \ln\left(\frac{\sigma_{ij}^{\mathrm{s}}}{\mu_i^{\mathrm{s}} \cdot \mu_j^{\mathrm{s}}} + 1 \right) \tag{6.15}$$

Modell: $\lambda_x^{\mathrm{s}} \rightarrow \lambda_x^{\mathrm{l}}$

Übergang in den Cepstralbereich: $\mathbf{x}^{\mathrm{c}} = \mathbf{A}_{\cos} \cdot \mathbf{x}^{\mathrm{l}}$

$$\mu^{\mathrm{c}} = \mathbf{A}_{\cos} \cdot \mu^{\mathrm{l}} \qquad \mathbf{C}^{\mathrm{c}} = \mathbf{A}_{\cos} \cdot \mathbf{C}^{\mathrm{l}} \cdot (\mathbf{A}_{\cos})^{\mathrm{T}} \tag{6.16}$$

Modelle: $\lambda_x^{\mathrm{l}} \rightarrow \lambda_x^{\mathrm{c}}$

Verwendbarkeit

Verwendet man die PMC zur einmaligen Anpassung an das additive Rauschen, so wird dadurch wie beim Training mit künstlich verrauschten Daten diese eine Störungsart in den Modellen berücksichtigt. Im Vergleich zum Training erfolgt die Anpassung etwas schneller, verwendet aber andererseits dafür etliche Näherungen. Prinzipiell ließe sich auch eine fortlaufende Anpassung realisieren (Online-PMC, siehe z.B. [Chiang 1997]), die für jede zu erkennende Anfrage durchgeführt wird. In diesem Fall würde man aus der aktuellen Äußerung über die Muster der Sprachpausen ein Geräuschmodell bestimmen und dieses jedes Mal mit dem akustischen Modell des Erkenners kombinieren.

Der praktischen Umsetzung der PMC zur Anpassung der akustischen Modelle eines Spracherkenners, wie er in dieser Arbeit verwendeten wird, stehen aber einige Dinge im Wege. Durch die Kanalkompensation (siehe Abschnitt 8.4.4) wird eine äußerungs- und sprecherabhängige Transformation hinzugefügt. Neben den cepstralen Merkmalen werden auch dynamische Merkmale (Delta-Koeffizienten) benutzt. Der erweiterte Merkmalsvektor (siehe Abschnitt 2.2 Merkmalsextraktion) wird ferner durch die LDA so transformiert, dass zwar die Diskriminanz der Klassen erhalten bleibt, sich die ursprünglichen Merkmalsvektoren aber aufgrund der Dimensionsreduktion nicht mehr rekonstruieren lassen. Durch bestimmte vereinfachende Annahmen und zusätzlichen Modellierungsaufwand [Gales 1995, Schless und Class 1997] lassen sich diese Probleme zwar mehr oder minder gut lösen, es bliebe dennoch die Aufgabe jede einzelne Gaußdichte aller phonetischen Klassen durch einen sehr aufwändigen Prozess zu bearbeiten. In dem hier zugrundeliegenden Basissystem würde dies bedeuten, die jeweils 32 Gaußdichten von 2.500 phonetischen Klassen, also insgesamt 80.000 Dichten, neu zu berechnen. Da hier nach einem Weg gesucht wird, einen Erkenner auf die augenblickliche Situation (für eine Äußerung) anzupassen, wäre dies sehr aufwändig und zeitintensiv. Wie weiter unten erläutert wird, bestehen auch Gemeinsamkeiten mit den hier untersuchten Trainings- und Adaptionsverfahren, mit denen ebenfalls an die Grenze des vertretbaren Aufwands gestoßen wurde.

Das Grundprinzip der Modellkombination soll dennoch als Baustein eines neuen Kompensationsverfahrens (siehe Kapitel 9: Modellbasierte Umgebungskompensation) dienen, bei dem sie nicht auf das akustische Modell des Spracherkenners angewendet werden muss und somit effizient umsetzbar bleibt.

6.4 Anpassung während der Erkennungsphase

Für die unmittelbare (fortlaufende) Anpassung an die jeweils aktuelle Äußerung stehen nur die dort extrahierten Muster zur Verfügung. Im Vergleich mit den Trainingsdaten sind dies also relativ wenige Beispiele, die meist nicht für eine gute Neuschätzung der Parameter ausreichen. Gelingt es eine annähernd genaue Zuordnung der Muster zu den entsprechenden phonetischen Modellen zu finden, lassen sich zumindest einige Modelle damit anpassen. Beim *Bayesian Learning* oder auch MAP-Adaption (Maximum-a-posteriori) [Gauvain und Lee 1994] werden die bestehenden Mittelwerte $\boldsymbol{\mu}_{s,m}$ der Gaußdichten mit den aus den neuen Daten geschätzen Werten $\hat{\boldsymbol{\mu}}_{s,m}$ interpoliert:

$$\tilde{\boldsymbol{\mu}}_{s,m} = \left(1-\alpha_{s,m}\right)\cdot \boldsymbol{\mu}_{s,m} + \alpha_{s,m}\cdot \hat{\boldsymbol{\mu}}_{s,m} \tag{6.17}$$

Der Interpolationsfaktor $\alpha_{s,m}$ bestimmt dabei wie zuverlässig die neu geschätzten Werte sind und wird in Abhängigkeit von der Anzahl der zur Bestimmung von $\hat{\boldsymbol{\mu}}_{s,m}$ verwendeten Muster $N_{s,m}$ und beispielsweise einem empirisch ermittelten Parameter η berechnet:

$$\alpha_{s,m} = \frac{N_{s,m}}{N_{s,m}+\eta} \tag{6.18}$$

Falls also keine Muster zur Neuschätzung vorliegen ($N_{s,m} = 0$), wird der alte Wert übernommen. Bei sehr vielen Mustern nähert sich der Mittelwert der Gaußdichte dem der Muster an. Ähnliche Überlegungen gelten auch für die Kovarianzmatrizen.

Um alle Modelle mit diesem Verfahren anzupassen, müssen jeweils ausreichend viele Daten aus einer gleichartigen Umgebung zur Verfügung stehen. Die Menge kann allerdings wesentlich geringer als die der zum Trainieren benutzen Sprachdaten sein. So eignet sich diese Methode beispielsweise zur Sprecheradaption. In [Crafa *et al.* 1998] wird sie dazu verwendet, den Rechenaufwand der Modellkombination dadurch zu verringern, dass weniger Daten generiert werden müssen.

Für den hier vorliegenden Fall zur Anpassung an die aktuelle Äußerung wurde ein anderes, ebenfalls verbreitetes Verfahren verwendet, das bekanntlich mit sehr wenigen Adaptionsdaten auskommt und im nun folgenden Abschnitt behandelt wird.

6.4.1 Maximum Likelihood Linear Regression (MLLR)

Grundprinzip der MLLR

Stehen nur wenige Daten aus einer neuen Umgebung zur Verfügung, können daraus nicht alle Parameter des akustischen Modells direkt neu bestimmt werden. Der bestehende Parametersatz kann jedoch durch eine **eingeschränkte Modifikation** so verändert werden, dass er besser zu den neuen Gegebenheiten passt, was mit einem entsprechenden **Kriterium** beurteilt werden muss.

Eine Modellannahme hilft bei Auswahl einer geeigneten Transformation. In Abschnitt 8.2 wird gezeigt, dass sich Kanal- und Sprechereigenschaften in einer zeitlich konstanten Verschiebung der cepstralen Merkmale bemerkbar machen. Um diesen Effekt auszugleichen, könnte man also alle Mittelwertvektoren der Gaußdichten durch einen gemeinsamen konstanten Verschiebungsvektor ändern [Sankar und Lee 1995]. Im etwas allgemeineren Fall nimmt man an, die Mittelwerte müssten linear transformiert werden. Es werden also veränderte Mittelwertvektoren $\tilde{\boldsymbol{\mu}}_{s,m}$ gesucht, die sich aus den ursprünglichen Modellparametern $\boldsymbol{\mu}_{s,m}$ durch eine Multiplikation mit einer konstanten Matrix **A** und Addition eines konstanten Verschiebungsvektors **b** ergeben.

$$\tilde{\boldsymbol{\mu}}_{s,m} = \mathbf{A} \cdot \boldsymbol{\mu}_{s,m} + \mathbf{b} \tag{6.19}$$

Anhand einiger Beispieldaten kann man diese Transformation so bestimmen, dass sich die Likelihood (siehe Abschnitt 3.3.4) der Daten maximiert. Diese *Maximum Likelihood Linear Regression* (**MLLR**) [Legetter und Woodland 1994, 1995] wurde ursprünglich zur Sprecheradaption verwendet. Die Umgebungen A und B (nach **Abbildung 6.1**, S. 84) stimmen zwar miteinander überein, das Modell wird aber auf den Spezialfall B_i (einen Sprecher) angepasst. Voraussetzung für dieses Verfahren ist eine Zuordnung der Muster zu den phonetischen Modellen und deren einzelnen Gaußdichten. Kennt man zumindest die Verschriftung einer Äußerung, kann mithilfe des Viterbi-Algorithmus diese Zuordnung gefunden werden. Ansonsten benötigt man für die Berechnung der Transformation nur noch die bisherigen Parameter der Gaußdichten (Mittelwerte und Kovarianzen) und die

zugeordneten Muster (Merkmalsvektoren). In [Gales und Woodland 1996] wird auch eine Möglichkeit beschrieben, ebenfalls die Kovarianzmatrizen anzupassen.

Statt einer gemeinsamen Transformationsmatrix für alle Modelle kann man diese (oder ihre Gaußdichten) auch in Regressionsklassen einteilen, für die jeweils eine eigene Transformation bestimmt wird. Sinnvollerweise verwendet man dazu einen Regressionsbaum, der die Modelle in immer kleinere Gruppen aufspaltet. Bei der letztendlichen Bestimmung der Regressionsklassen verzweigt man nur soweit in diesen Baum, dass für die Gruppen noch eine Mindestanzahl von Beispielen vorhanden ist. Für die unten vorgestellten Experimente wurde der Regressionsbaum über die Gaußdichten aller Modell mithilfe des *„Neural-Gas"*-Verfahrens bestimmt ([Martinez *et al.* 1993], siehe auch Seite 66). Für jede Verzweigung wurden dabei die noch verbleibenden Gaußdichten anhand ihrer Mittelwerte in zwei Ballungen zerlegt. Die minimale Musteranzahl zur Bildung einer Regressionsklasse bei der Adaption betrug 250, was 2,5 Sekunden Sprachdaten entspricht.

Anwendung der MLLR

Ist man bereit, den Sprecher vor der eigentlichen Nutzung eines Spracherkenners einige vorgegebene Äußerungen sprechen zu lassen, erhält man dadurch eine Adaptionsmenge inklusive Verschriftung. Es handelt sich hierbei um eine **überwachte Adaption**, denn man kann von der Korrektheit der Verschriftung ausgehen. Während dieses Vorgehen bei Diktieranwendungen gängige Praxis ist, kann man bei Informationssystemen meist nicht diesen Kompromiss eingehen, da die Benutzer zum Teil nur wenige Anfragen stellen wollen.

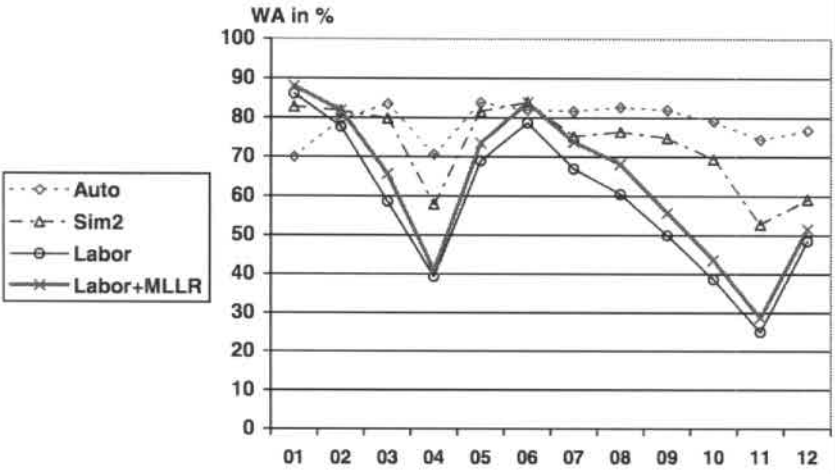
Als Näherung ersetzt man bei der **unüberwachten Adaption** die korrekte Verschriftung mit einer ersten Hypothese des noch nicht angepassten Erkenners. Da diese fehlerhaft sein kann, kommt es teilweise vor, dass durch die Adaption das Ergebnis sogar noch weiter verschlechtert wird. Werden jedoch die Muster einigermaßen stimmig den phonetischen Klassen zugeordnet, führt die gemeinsame Transformation aller Phonemmodelle zu einer besseren Übereinstimmung des akustischen Modells mit den aktuellen Sprachmustern.

Auf eine iterative Anwendung wurde im vorliegenden Fall verzichtet. Durch die angepassten Modelle könnten zwar damit bessere Verschriftungen (und damit bessere Zuordnungen) gefunden werden, um abermals zu adaptieren, die jeweils nötigen Erkennungsläufe stellen aber einen großen Rechenaufwand dar. Das Vorgehen zur Erkennung einer Äußerung mit MLLR lautet also folgendermaßen:

Unüberwachte Adaption mit MLLR für eine Äußerung:

1. Erste Erkennung mit dem noch nicht angepassten Modellen:
   $\Rightarrow$ Hypothese H_0 (Verschriftung).
2. Zuordnung der Muster zu den Modellen mittels H_0 durch den Viterbi-Algorithmus.
3. Adaption der Modelle:
   $\tilde{\boldsymbol{\mu}}_{s,m} = \mathbf{A} \cdot \boldsymbol{\mu}_{s,m} + \mathbf{b}$
4. Erneute Erkennung mit den adaptierten Modellen.

Abbildung 6.6 zeigt die Ergebnisse für diese Art der Anpassung im Vergleich mit den Referenzsystemen. Die erreichten Fehlerreduktionen $ER_{relativ}$ gegenüber dem Basissystem reichen von 24% für das stehende Fahrzeug (Kategorie 06) bis zu 3% bei starkem Lüftergeräusch (Kategorie 04). Die relative Fehlerreduktion für den ungestörten Fall (Kategorie 01) beträgt 14 %. Dieser Wert wurde auch auf dem gleichen Testset mit dem Nahbesprechungsmikrofon erreicht, bei dem sich die Wortakkuratheit von 86,9% auf 88,7% steigerte.



Kategorie	01	02	03	04	05	06	07	08	09	10	11	12
System	aus	Blinker	Fenster	stark	mittel	0km/h	25km/h	50km/h	75km/h	100km/h	125km/h	beacht.
Basiserkenner (WA in %)	86,1	77,7	58,6	39,4	68,9	78,7	66,9	60,5	50,0	38,7	25,0	48,5
mit MLLR (WA in %)	88,0	81,9	65,6	41,1	73,5	83,9	73,8	68,1	55,8	43,6	28,7	51,7
ER relativ	14%	19%	17%	3%	15%	24%	21%	19%	12%	8%	5%	6%

Abbildung 6.6: Verbesserung des Laborerkenners durch unüberwachte Adaption.

Die Adaption mit MLLR bringt hier also durchweg Verbesserungen des Basiserkenners. Je stärker sich aber die Bedingung von der ungestörten Trainingsumgebung unterscheidet und damit zu einer schlechten ersten Hypothese führt, desto weniger effektiv arbeitet die Anpassung. Für diese Fälle weichen die absoluten Erkennungsraten deutlich von den beiden Referenzsystemen mit verrauschten Trainingsdaten ab und stellen somit keine geeignete Alternative dar. Zwei Schwächen, nämlich die geringe Adaptionsdatenmenge und die unzuverlässige Verschriftung werden durch den im nächsten Abschnitt vorgestellten Ansatz behoben.

6.4.2 Online Simulated Training Data (OSTD)

Grundprinzip

Mit einem neuen Ansatz soll nun auf der Basis eines Adaptionsverfahrens (hier gezeigt für die MLLR) der Nachteil beseitigt werden, der sich aus den sonst zugrundeliegenden, fehlerhaften Hypothesen ergibt. Dazu wird ähnlich wie bei den simulierten Trainingsdaten die neue Umgebung mit vorhandenen Trainingsdaten nachgestellt. Da wir es hier mit einer konkreteren Situation zu tun haben, kann man jedoch gegenüber dem Training noch einen Schritt weiter gehen. Aus der vorliegenden Äußerung werden sowohl Rauschsegmente entnommen als auch Sprechereigenschaften geschätzt. Dann werden Äußerungen ähnlicher Sprecher aus der Trainingsmenge mit dem aktuellen Rauschen kombiniert. Das heißt, sowohl das aktuelle Hintergrundrauschen wie auch der Sprecher werden simuliert. Die Zuordnung der Muster zu den phonetischen Klassen kann auf den noch unverrauschten Trainingsäußerungen mittels der Referenztexte vorberechnet werden. Damit liegt am Ende eine optimale Zuordnung von gestörten Merkmalsvektoren zu den zugehörigen Klassen vor.

Die Vorteile gegenüber der unüberwachten MLLR sind also:

- Korrekte Verschriftung des Adaptionsmaterials.
- Verwendung der ungestörten Adaptionssätze bei der Zuordnung, aber dennoch Bestimmung der Anpassung mit verrauschten Mustern.
- (Fast) beliebige Menge an Adaptionsmaterial, das die Situation der aktuellen Aufnahme bezüglich Sprecher und Rauschen widerspiegelt.

Bevor auf die Realisierungsdetails eingegangen wird, hier die schematische Zusammenfassung dieser vom Autor OSTD benannten Anpassung (von engl.: *online simulated training data*) zur Erkennung einer Aufnahme. Die Verwendung der Sprechereigenschaften sowie ein abschließender unüberwachter Adaptionsschritt sind optional und daher in Klammern angegeben:

OSTD-Adaption für eine Äußerung:

1. Extraktion von reinen Geräuschsegmenten aus der aktuellen Aufnahme (und Schätzen von Sprechereigenschaften)
2. Selektion einer Menge von Äußerungen aus der Trainingsmenge (fest oder aufgrund ähnlicher Sprechereigenschaften). Die Zuordnung der Muster zu phonetischen Klassen liegt bereits vor.
3. Kombination der selektierten Trainingsäußerungen mit dem extrahierten Rauschen
4. Adaption (z.B. MLLR) der Modelle mit den simulierten Trainingsdaten
5. Erkennung der Aufnahme mit den modifizierten Modellen
6. (weitere unüberwachte Adaption aufgrund der im vorigen Schritt gewonnenen Hypothese und basierend auf dem bereits angepassten Modell)

Umsetzung

Um Proben des additiven Rauschens zu erhalten, werden mithilfe eines Sprach-Pause-Detektors Pausensegmente innerhalb der Äußerung bestimmt und dort komplette Analyserahmen aus dem Zeitsignal entnommen. Diese Analyserahmen werden durch mehrfaches Hintereinandersetzen der Segmente auf die Länge der selektierten Trainingsäußerungen gebracht und können nun zu den dort entnommenen Rahmen addiert werden. Es wird hier nicht mit den Zeitsignalen, sondern mit den entnommenen Analyserahmen gearbeitet, um sicherzustellen, dass beim Aneinanderfügen der Pausensegmente innerhalb eines Analyserahmens keine unterschiedlichen Rauschsegmente aufeinanderstoßen. Merkmale, die zum Teil mithilfe unterschiedlicher Rauschsegmenten berechnet werden (z.B. Delta-Koeffizienten), können jedoch auch so nicht völlig ausgeschlossen werden.

Als einfache Eigenschaft zur Auswahl der Sprecher könnte z.B. das Geschlecht des Testsprechers verwendet werden. Eine genauere Einteilung der Sprecher erlaubt jedoch die Schätzung der Vokaltraktlänge über den sogenannten Verzerrungsfaktor der Vokaltraktlängen-Normierung VTLN (siehe Kapitel 7). Der Verzerrungsfaktor wurde für die hier aufgeführten Experimente mit dem Maximum-Likelihood-Ansatz gewonnen und liegt als quantisierter Wert vor. Da die gestörten Testaufnahmen einer zuverlässigen Schätzung dieses Faktors nicht gerade förderlich sind, wurde neben Trainingssprechern mit demselben Verzerrungsfaktor auch solche mit benachbarten Quantisierungswerten selektiert.

Da Trainingsdaten in ausreichender Menge vorliegen, ergibt sich eine Beschränkung der zur Adaption verwendeten Äußerungen nur durch die Zeitvorgabe zur Erkennung der Testäußerung. Da dies natürlich auch von der Implementierung und der Prozessorleistung abhängt, wurde die Anzahl einigermaßen großzügig auf 20 festgesetzt.

Versuche mit künstlich verrauschten Testdaten

In einem ersten Experiment wurde zunächst Fahrgeräusch (100 km/h) gewichtet zu den Testaufnahmen der ungestörten Kategorie 01 zugemischt. Wie aus den SNR-Werten in Abschnitt 4.1.2 zu sehen ist, entspricht der Mischungsfaktor 0,5 etwa einer mittelstarken Störung im Fahrzeug. SNR-Werte wie für den Mischungsfaktor 1,0 werden unter realistischen Bedingungen nie erreicht, da sich die Sprecher in der Lautstärke anpassen.

Tabelle 6.2 zeigt die Ergebnisse (Sennheiser, Kategorie 01) für den Basiserkenner und einer abgeleiteten Version, bei der die Modelle überwacht mit MLLR einmalig adaptiert wurden. Die fest selektierte Adaptionsmenge besteht aus 20 ungestörten Sprachaufnahmen verschiedener Sprecher und wurde auch bei der „festen" OSTD verwendet (s.u.). Wird der Basiserkenner einmalig mit diesen ungestörten Aufnahmen adaptiert, liefert dies ähnliche Resultate für den Testfall ohne und mit zugemischtem Rauschen. Die Erkennungsrate bricht für beide Erkenner durch das addierte Störgeräusch von etwa 87 % auf 60 % ein.

Tabelle 6.3 zeigt Ergebnisse für äußerungsbasierte Anpassungen des Basiserkenners. Mit der unüberwachten MLLR, basierend auf der ersten Erkennungshypothese der gestörten Äußerung (also mit etwa 60 % WA), kann das Ergebnis auf 66,2 % verbessert werden. Verwendet man hingegen die 20 festen Äußerungen mittels OSTD werden 81,2 % erreicht. Mit einer Auswahl des Adaptionsmaterials (in gleicher Menge) von bestimmten Sprechern, durch Vergleich mit dem automatisch bestimmten Verzerrungsfaktor des Testsprechers,

ergeben sich nur noch leichte, zusätzliche Gewinne. Insgesamt beträgt so die relative Fehlerreduktion gegenüber dem nicht angepassten Basissystem über 50 %.

Tabelle 6.2: Einmalige überwachte MLLR-Anpassung.

WA in % Erkenner: / Testbedingung:	**Basissystem**	**Basissystem mit einmaliger, überwachter MLLR auf fester, ungestörter Adaptionsmenge**
Ohne Rauschen		
Normal	86,9	87,2
Fahrgeräusch zugemischt		
Normal	60,5	59,1

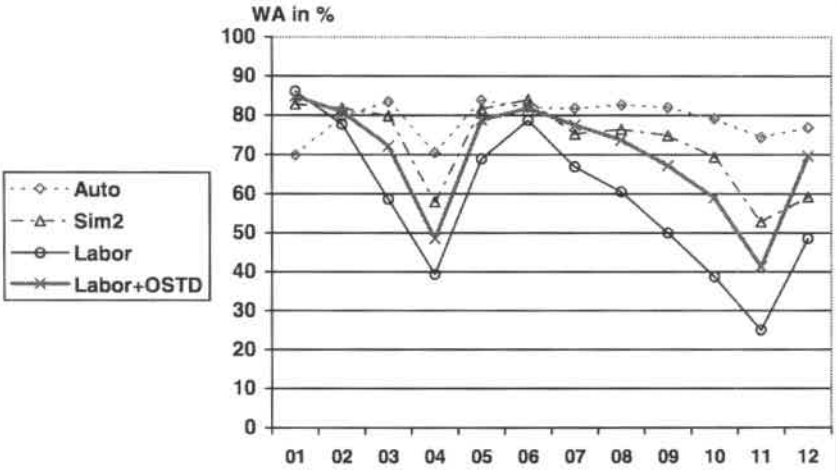
Tabelle 6.3: Vergleich äußerungsbasierter MLLR- und OSTD-Anpassung auf künstlich verrauschten Testdaten.

Erkenner: / Testbedingung:	**Basissystem** WA in %	relative Fehlerreduktion
Fahrgeräusch zugemischt		
Normal	60,5	-
unüberwachte MLLR	66,2	14 %
feste OSTD	81,2	52 %
Sprecher-OSTD	81,7	54 %

Versuche mit Autosprachdaten

Diese im Vergleich mit der unüberwachten MLLR außerordentlich guten Verbesserungen konnten allerdings nicht in voller Höhe auf reale Daten übertragen werden. Aus **Abbildung 6.7** entnehmen wir, dass die OSTD-Adaption ohne Sprecherselektion dennoch Fehlerreduktionen bis zu 40% erzielt und im Schnitt deutlich besser als die unüberwachte Adaption mit MLLR abschneidet (vergleiche **Abbildung 6.6**, S. 98).

Für die ungestörte Kategorie 01 verschlechtern sich die Ergebnisse hingegen. Hier liegt die Leistung des geschätzten Rauschens ja auch ungefähr auf dem Niveau des Restrauschens der Trainingsaufnahmen und wird durch die Addition bei den simulierten Daten verstärkt. Das simulierte Rauschen entspricht somit nicht mehr den Gegebenheiten. Für Aufnahmen mit stärkerem Hintergrundrauschen kann man dagegen den Einfluss des Restrauschen in den Trainingsdaten vernachlässigen.

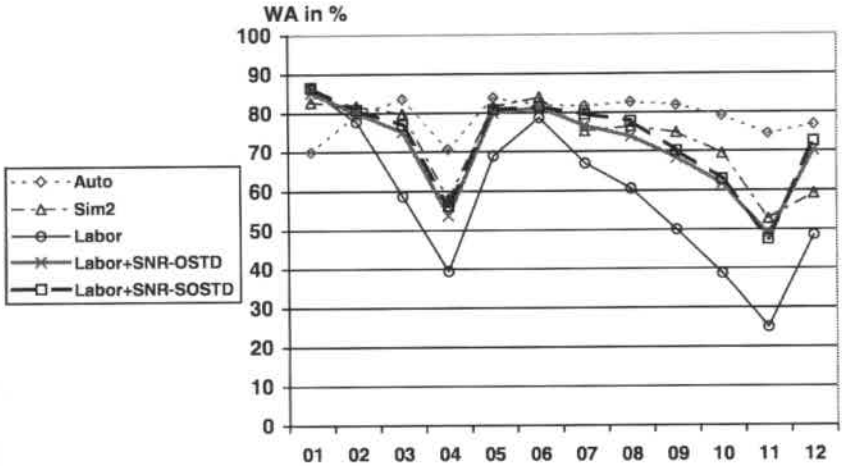


Kategorie	01	02	03	04	05	06	07	08	09	10	11	12
System	aus	Blinker	Fenster	Lstark	Lmittel	0km/h	25km/h	50km/h	75km/h	100km/h	125km/h	beschl.
Basiserkenner (WA in %)	86,1	77,7	58,6	39,4	68,9	78,7	66,9	60,5	50,0	38,7	25,0	48,5
mit OSTD (WA in %)	84,8	80,9	72,2	48,5	78,8	81,8	77,5	73,7	67,2	59,0	41,3	69,6
ER relativ	-9%	14%	33%	15%	32%	15%	32%	33%	34%	33%	22%	41%

Abbildung 6.7: Ergebnisse der OSTD-Adaption ohne Sprecherselektion.

Schon beim Vergleich der SNR-Werte fiel auf, dass das Zumischen von Rauschen nicht zur selben Intensität der Störung führt wie unter realistischen Bedingungen. Wird also das aus der Testaufnahme entnommene Geräusch zu den ungestörten Trainingsdaten gemischt, erhält man damit nicht die gleichen Bedingungen wie für die Testaufnahme. Daher wurde für weitere Versuche vor dem Zumischen eine SNR-Normalisierung durchgeführt. Sie sorgt dafür, dass der SNR der simulierten Daten dem SNR der Testaufnahme entspricht.

Tatsächlich konnten so die Ergebnisse nochmals gesteigert werden (SNR-OSTD in **Abbildung 6.8**). Führte man nun noch zusätzlich die Sprecherselektion durch (SNR-SOSTD) wurden durchgängig positive Fehlerreduktionen erreicht, die bis zu 44 % betragen.



Kategorie	01	02	03	04	05	06	07	08	09	10	11	12
System	aus	Blinker	Fenster	Lstark	Lmittel	0km/h	25km/h	50km/h	75km/h	100km/h	125km/h	beschl.
Basiserkenner (WA in %)	86,1	77,7	58,6	39,4	68,9	78,7	66,9	60,5	50,0	38,7	25,0	48,5
+ SNR-OSTD (WA in %)	85,2	79,9	75,1	53,7	79,9	81,3	76,6	73,8	68,3	61,6	48,6	70,1
ER relativ	-6%	10%	40%	24%	35%	12%	29%	34%	37%	37%	31%	42%

Kategorie	01	02	03	04	05	06	07	08	09	10	11	12
System	aus	Blinker	Fenster	Lstark	Lmittel	0km/h	25km/h	50km/h	75km/h	100km/h	125km/h	beschl.
Basiserkenner (WA in %)	86,1	77,7	58,6	39,4	68,9	78,7	66,9	60,5	50,0	38,7	25,0	48,5
+ SNR-SOSTD (WA in %)	86,6	80,8	76,9	56,2	80,9	81,5	79,6	77,8	70,1	62,8	47,5	72,5
ER relativ	4%	14%	44%	28%	39%	13%	38%	44%	40%	39%	30%	47%

Abbildung 6.8: Ergebnisse der SNR-normalisierten OSTD-Adaption ohne und mit Sprecherselektion.

Mit der SNR-normalisierten OSTD-Adaption mit Sprecherselektion konnte also eine Anpassung des Basiserkenners auf die augenblickliche Situation erreicht werden, die im Vergleich mit der unüberwachten MLLR deutlich bessere Ergebnisse liefert. Diese sind durchaus vergleichbar mit Resultaten, die der neu auf künstlich verrauschten Daten trainierte Erkenner „Sim2" erreicht. Gegenüber der unüberwachten MLLR erspart man sich dabei auch die zeitraubende Erkennung einer ersten Hypothese. Die äußerungsweise Anpassung des kompletten akustischen Modells des Erkenners mit seinen 2.500 Gaußmixturen nimmt jedoch immer noch sehr viel Zeit in Anspruch, die eine Erkennung in beinahe Echtzeit nicht zulassen.

6.5 Zusammenfassung

Trotz der deutlichen Verbesserungen des Laborerkenners durch die äußerungsbasierten Adaptionsmaßnahmen (man bedenke, dass im Vorfeld keinerlei Daten aus der Autoumgebung verwendet wurden), ergeben sich für einige Kategorien noch sehr unbefriedigende Ergebnisse, die zum Teil durch den Erkenner „Sim2" übertroffen werden. Andererseits könnten theoretisch auch für die OSTD beliebig viele Daten benutzt werden, was aber den ohnehin für die Echtzeiterkennung zu hohen Rechenaufwand weiter erhöhen würde. Stehen für eine neue Umgebung ähnlich viele Geräusch- und Sprachdaten zur Verfügung wie für „Sim2", könnte man die MLLR bei der OSTD durch eine Neuschätzung der Daten ersetzen und würde zu den Ergebnissen von „Sim2" gelangen.

Vergleich der OSTD mit der Modellkombination

Wollte man die Modellkombination (PMC, Abschnitt 6.3) für die Anpassung eines Laborerkenners auf einen einzelnen, verrauschten Satz anwenden, würde man zunächst ein einfaches Geräuschmodell λ_n auf Grundlage der als Sprachpausen detektierten Merkmalsvektoren herleiten. Nehmen wir ferner an, die Modellkombination wird mit einem datengetriebenen Ansatz durchgeführt. Der Unterschied zur OSTD-Adaption (oder simulierten Trainingsdaten) besteht dann darin, dass keine realen Beispiele für Sprache und Geräusch kombiniert werden, sondern zufällig generierte aus einem jeweils vorliegenden Modell. Für die additive Störung, für die sonst nur wenige unterschiedliche Beispiele existieren, mag dies ein Vorteil sein. Bei der Erzeugung der Sprachdaten müssen jedoch etliche Näherungen verwendet werden (wenn sie sich überhaupt wieder aus den Modellen rückrechnen lassen), so dass Merkmalsvektoren „echter" Daten wohl geeigneter erscheinen. Diese Daten werden bei der PMC wie beim Training mit simulierten Daten zur Neuschätzung der Modellparameter verwendet, wofür entsprechend viele Muster generiert werden müssen. Bei der oben vorgestellten OSTD-Adaption wurde hingegen angenommen, dass dafür die Menge der Beispiele nicht ausreichend ist. Daher wurde dort die Anpassung mittels der MLLR auf eine lineare Transformation der bestehenden Parameter beschränkt. Würde man den Aufwand ähnlich hochtreiben wie für eine datengetriebene Modellkombination und dann zu einer Neubestimmung der Parameter übergehen, ist wohl mit mindestens gleichwertigen Ergebnissen zu rechnen. Der Vorteil der Modellkombination liegt dann eher in der kompakteren Darstellung des Trainingmaterials für ungestörte Sprache. Ein klarer Vorteil der OSTD liegt hingegen in der Möglichkeit der Sprecherselektion.

7 Sprechernormierung

Die Entdeckung,
dass es so einfach nicht ist wie man gedacht hat,
ist als Gewinn anzusehen.
- Carl Friedrich von Weizsäcker -

Kapitelübersicht: Dieses Kapitel befasst sich mit der Kompensation der Änderungen im Sprachsignal, die durch wechselnde Sprecher entstehen. Auch ein und derselbe Sprecher wird niemals Äußerungen mit gleichem Inhalt völlig identisch realisieren können. Diese Variationen sollen hier aber nicht betrachtet werden, vielmehr Eigenschaften, die für einen Sprecher typisch sind und sich stark in den vom Erkenner benutzten Merkmalen niederschlagen. Für die „Sprecher"-Erkennung sind gerade diese Eigenschaften besonders interessant und hervorzuheben. Im Falle der Spracherkennung wirken sie sich aber störend aus. Sie vergrößern die Varianz der phonetischen Klassen im Merkmalsraum, so dass diese sich schwerer unterscheiden lassen.

Ausgehend von einem Modell der Spracherzeugung und des Vokaltrakts (Abschnitt 7.1) werden zunächst allgemeine Überlegungen zu Sprechereigenschaften im Sprachsignal angestellt (Abschnitt 7.2). Die folgenden Abschnitte konzentrieren sich dann auf die Frequenzverzerrung zur Sprechernormierung (auch Vokaltraktlängennormierung genannt). Nach einer Einführung und Darstellung der verschiedenen Möglichkeiten (Abschnitt 7.3) wird die in dieser Arbeit umgesetzte ML-VTLN beschrieben (Abschnitt 7.4). Durch ein anderes vom Autor vorgeschlagenes Optimierungskriterium werden die Trainingseigenschaften dieses modellbasierten Ansatzes verbessert (Abschnitt 7.5). Im letzten Abschnitt (7.6) werden die Verfahren experimentell verglichen und ihre Verwendung für den mobilen Einsatz diskutiert.

7.1 Sprechermodell

Modelle der menschlichen Sprachproduktion (Artikulationsorgane) und der Sprachrezeption (Gehör) dienen bei der Spracherkennung als wichtige Basis zur Analyse von Sprachsignalen. Eigenschaften der menschlichen Lautwahrnehmung werden beispielsweise bei der Merkmalsextraktion nachempfunden (Beispiel: *Mel*-Skala). Auch wenn es etwas schwer fällt Eigenschaften der Sprecher als Störung zu bezeichnen, soll nun mit einem einfachen Sprechermodell ein Zusammenhang beschrieben werden, der eine maschinelle Erkennung erschwert. Bedingt durch unterschiedliche Anatomien der jeweiligen Sprecher ergeben sich nämlich durch Sprecherwechsel starke Schwankungen der zu erkennenden Sprachsignale.

Ein einfaches schematisches Modell der Sprachproduktion zeigt **Abbildung 7.1**. Auf weitere Details wird an dieser Stelle verzichtet und stellvertretend für tiefergehende Literatur an [Rabiner und Schafer 1978] verwiesen. Nur soviel sei bemerkt: Für die Spracherkennung von Interesse sind hauptsächlich die Parameter des Vokaltraktmodells, die für die Lautformung verantwortlich sind. Sie drücken sich bei stimmhaften Lauten in Form von Resonanzstellen im Spektrum aus (Formanten). Die Anregungsfunktionen spielen dagegen meist eine untergeordnete Rolle. Die durch die Stimmbänder hervorgerufene periodische Anregungsfunktion ist stark sprecherabhängig (Frauen haben höhere Stimmlagen als Männer). Da die Anregungsfunktion gewissermaßen nur als Träger der Lautinformation dient, wird sie bei der Merkmalsextraktion herausgefiltert.

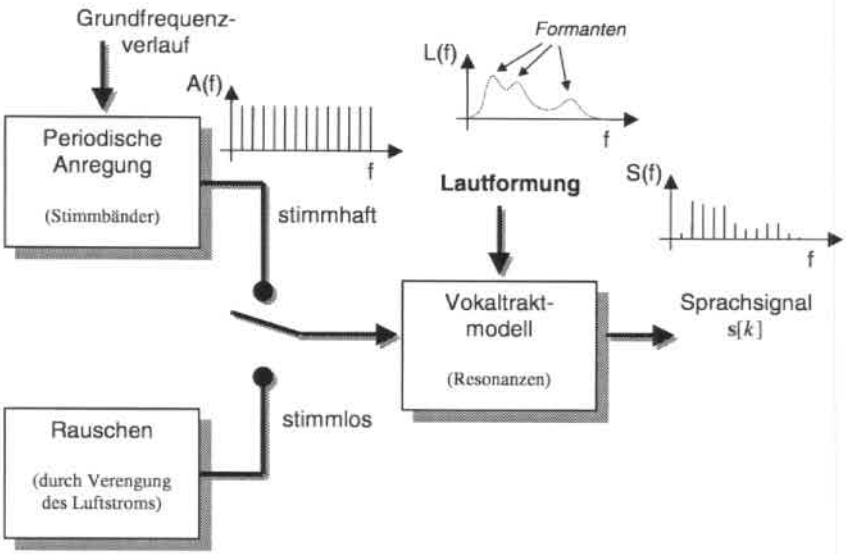


Abbildung 7.1: Schematisches Modell der Sprachproduktion.

Das klärt die prinzipielle Entstehung der Sprachlaute. Wie hängen diese jedoch vom Sprecher ab? **Abbildung 7.2** zeigt einen Schnitt durch den menschlichen Vokaltrakt (basierend auf [Kohler 1977]). Bedingt durch die unregelmäßige Form und andere Faktoren lassen sich hierfür beliebig komplizierte Modelle konstruieren (in [Rabiner und Schafer 1978] finden sich einige Beispiele dafür). Um einen groben Zusammenhang des Sprachsignals von verschiedenen Sprechern zu erhalten, reicht es aber den gesamten Vokaltrakt als einfache Röhre mit einem offenen Ende zu betrachten (**Abbildung 7.3**). Diese als verlustfrei angenommene Röhre besitzt Resonanzstellen in regelmäßigen Abständen:

$$f_n = \frac{(2n+1)\cdot c}{4L} \qquad n = 0,1,2,\ldots \tag{7.1}$$

Die Schallgeschwindigkeit c kann dabei für gewöhnlich als konstant betrachtet werden (indem man seinen Vokaltrakt beispielsweise mit Helium füllt, kann man diese Abhängigkeit jedoch sehr effektvoll zeigen). Die Resonanzfrequenzen hängen also in erster Linie von der Länge L des Vokaltraktes ab.

Die Länge wiederum ist stark sprecherabhängig. Bei Männern liegt sie im Mittel bei etwa 18 cm, bei Frauen hingegen bei durchschnittlich 13 cm. Der Vergleich von Formanten von Männern und Frauen bestätigt tendenziell den umgekehrt proportionalen Zusammenhang aus Gleichung (7.1). Bei anderen Lauten (wie Frikativen, Plosiven) unterscheiden sich hingegen die Art und der Ort der Schallerzeugung sehr stark von dem einfachen Röhrenmodell. Daher ist dieser Zusammenhang von Resonanzfrequenz und Vokaltraktlänge generell auf stimmhafte Laute (Anregung durch die Stimmbänder) beschränkt.

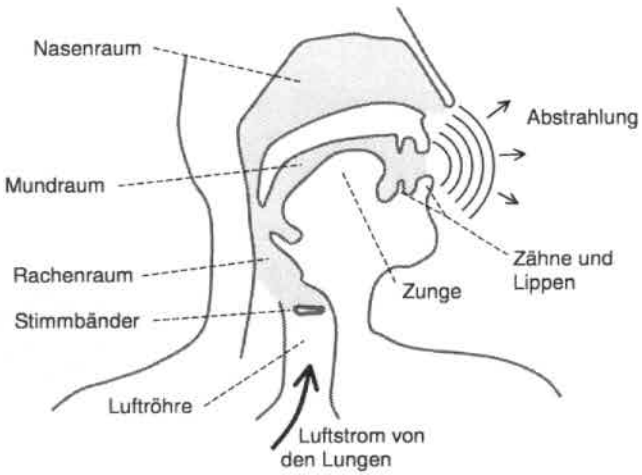


Abbildung 7.2: Querschnitt des menschlichen Vokaltraktes.

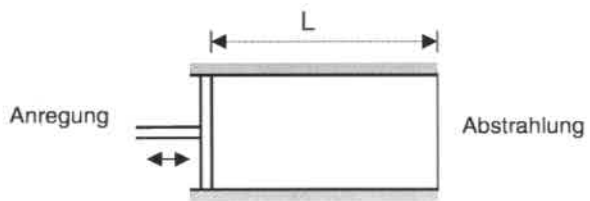


Abbildung 7.3: Röhrenmodell des Vokaltraktes mit der Länge L.

7.2 Sprechereigenschaften

Sprechgeschwindigkeit

Die Sprechgeschwindigkeit ist für starke Variationen der Sprachrealisierung verantwortlich. Es gibt zwar tendenziell schnelle und langsame Sprecher, die Sprechgeschwindigkeit jedoch ist keine feste Eigenschaft eines Sprechers, da die Geschwindigkeit der geäußerten Lautfolgen erheblich von Satz zu Satz schwanken kann. Auch innerhalb einer Äußerung findet man große Unterschiede, was gerade für spontane Sprache sehr typisch ist. Obwohl diese Eigenschaft also thematisch nicht unbedingt in dieses Kapitel gehört, soll sie hier dennoch kurz gestreift werden, da sie ansonsten nicht weiter behandelt wird (für Arbeiten, die sich mit diesem Thema eingehender beschäftigen sei an [Siegler und Stern 1995, Morgan *et al.* 1997, Pfau *et al.* 1999] verwiesen).

Eine zeitliche Zuordnung der unterschiedlich schnellen Lautfolgen wird bereits durch die Verwendung von HMMs bewerkstelligt. Natürlich ändert sich durch eine schnellere oder langsamere Realisierung nicht nur die Anzahl der Merkmalsvektoren für einen Laut (HMM-Zustand), sondern auch deren Erscheinungsbild. Manche Phoneme eines schnell gesprochenen Wortes werden gar völlig verschluckt oder „verschmieren" mit den benachbarten Phonemen (Verschleifung). Diese Effekte, sofern sie in den Trainingsdaten zu finden waren, können teilweise auch ohne explizite Modellierung durch das akustische Modell des Erkenners berücksichtigt werden. Im vorliegenden Fall basiert der verwendete Erkenner auf kontextabhängigen phonetischen Einheiten, die mithilfe von spontansprachlichen Trainingsdaten geballt wurden (siehe Abschnitt 4.2). Häufige und hinreichend verschiedene Verschleifungen eines Phonems werden also bereits abhängig vom phonetischen Kontext unterschieden. Einen zusätzlichen Gewinn wird man nur erzielen, wenn für einen betrachteten Kontext noch abweichende Varianten durch zusätzliche Information vorhergesagt werden können. Dies kann eventuell aufgrund der Sprechgeschwindigkeit geschehen.

Ein etwas anderer Weg wurde im Laufe dieser Arbeit eingeschlagen. Mit einer Schätzung der mittleren Sprechgeschwindigkeit einer zu erkennenden Äußerung wurde die Analyseschrittweite angepasst. Das heißt, die spektralen Merkmale ändern sich dadurch prinzipiell nicht, aber die Häufigkeit der Stichproben. Dadurch sollten vor allem die dynamischen Merkmale bezüglich der Sprechgeschwindigkeit normiert werden. Dies führte aber nur zu minimalen Verbesserungen und wurde daher nicht weiter verfolgt.

Langzeitspektrum

Eine etwas typischere Sprechereigenschaft ist das Langzeitspektrum, das man näherungsweise über die Mittelung der Kurzzeitspektren (Gleichung (2.7), S. 19) berechnen könnte. Bedingt durch die Form und Größe des Vokaltraktes, Sprache, Sprechweise und andere Faktoren ergeben sich für einen Sprecher typische Frequenzverteilungen. Durch die im nächsten Kapitel behandelten Kanalkompensationsverfahren wird dieser *„Sprecherkanal"* ausgeglichen. In der Praxis ist es sowieso schwierig anhand des Sprachsignals zwischen diesem *Sprecherkanal* und dem eigentlichen *Übertragungskanal* zu unterscheiden. Diese Unterscheidung ist eher für die Sprechererkennung relevant und braucht hier nicht gesondert behandelt zu werden.

Vokaltrakt

Neben dem Langzeitspektrum wirkt sich die Form und Größe des Vokaltraktes insbesondere auf die **Resonanzstellen (Formanten)** verschiedener Laute aus. Diese sind natürlich sehr wichtig zur Unterscheidung der phonetischen Klassen, ihre Ausprägung schwangt aber auch stark von Sprecher zu Sprecher. Es wurde daher versucht *robustere Merkmale* (vergleiche Abschnitt 5.4) zu finden, die eher die Verhältnisse der Resonanzstellen zueinander berücksichtigen und nicht so sehr ihre absolute Lage im Spektrum [Miller 1989]. Diese Ansätze konnten sich aber im Gegensatz zu entsprechenden *Normierungsverfahren*, die auch in der vorliegenden Arbeit untersucht wurden, bisher nicht durchsetzen. Grundlage beider Ansätze ist in den meisten Fällen die Annahme, dass sich die Lage der Formanten F_i umgekehrt proportional zur Länge des Vokaltraktes L_{VT} verhält (nach Gleichung (7.1)).

Lineares Vokaltraktmodell: $$F_i \sim \frac{1}{L_{VT}} \tag{7.2}$$

Daher leitet sich auch der häufig benutzte Name Vokaltraktlängen-Normierung (VTLN) ab, der schon bei [Wakita 1977] verwendet wurde, um isoliert gesprochene Vokale durch Normierung besser zu klassifizieren. Zwar ist das Ziel dieser Ansätze nicht unbedingt die exakte Normierung von L_{VT}, aber durch die Beziehung zu den Resonanzstellen im Spektrum trifft es die Sache.

7.3 Vokaltraktlängen-Normierung (VTLN)

7.3.1 Grundprobleme der VTLN

Wie bei den anderen Kompensationsverfahren liegt auch bei der VTLN eine Modellannahme der Störung zugrunde. Aus dieser kann eine *Normierungsfunktion* abgeleitet werden, deren *Parameter* aus dem vorliegenden Sprachsignal gewonnen werden. Nimmt man die oben angegebene einfache Beziehung als Grundlage für das Modell, hat man es mit nur einem Parameter zu tun: der Länge L_{VT}. Ist diese bekannt, muss das Spektrum linear entlang der Frequenzachse so verzerrt werden, dass sich die Formanten eines Standardsprechers ergeben. Damit hat man die folgenden beiden Probleme zu lösen:

1. Wie wird das Sprachsignal transformiert,
   damit sich eine lineare Verzerrung der Frequenzachse ergibt?
2. Wie findet man einen Parameter,
   der zu einer optimalen Verzerrung führt?

In Anlehnung an die Literatur zu dieser Thematik wird hier statt der Vokaltraktlänge eines Sprechers a der allgemeinere Begriff Verzerrungsfaktor (engl.: *warping factor*) verwendet. Dieser ist hier als Reziprok der Vokaltraktlänge definiert. Für männliche Sprecher mit meist längerem Vokaltrakt und Formanten bei niedrigeren Frequenzen ergibt sich also ein kleiner Verzerrungsfaktor α_a, für weibliche Sprecher entsprechend umgekehrt. Als Standardsprecher gelte ein Sprecher mit mittleren Werten und $\alpha_a=1$.

7.3.2 Verzerrung der Frequenzachse

Eine Verzerrung der Frequenzachse kann praktisch in fast allen Schritten der üblichen Merkmalsextraktion durchgeführt werden. Durch Abtastung im **Zeitbereich** mit unterschiedlichen Abtastraten f_A ergibt sich nach der Kurzzeit-Spektralanalyse ein Frequenzbereich zwischen 0 Hz und der halben Abtastrate. Wird die Abtastrate entsprechend dem Verzerrungsfaktor gewählt und die Anzahl der Analysewerte im Fenster gleich belassen, erreicht man eine lineare Frequenzanpassung. Dadurch ändert sich allerdings auch die zeitliche Ausdehnung des Analysefensters. Die Schrittweite der Fenster in Sekunden sollte trotz unterschiedlicher Abtastraten wegen der zeitlichen Modellierung im Erkenner konstant gehalten werden. Es ergeben sich allerdings damit unterschiedliche Überdeckungen der Fenster für verschiedene Verzerrungen, was von Nachteil ist, da sich dadurch die dynamischen Merkmale stark verändern. Trotz einiger Vorzüge ist diese Methode nicht üblich, denn meist liegen die Sprachdaten in einem einheitlichen Format mit konstanter Abtastrate vor. Ansonsten müsste der Verzerrungsfaktor schon vor der Digitalisierung bekannt sein. Eine Neuabtastung bereits digitalisierter Daten führt durch das Abtasttheorem (Bandbegrenzung auf halbe Abtastrate) auf ähnliche Probleme wie eine Transformation im Frequenzbereich (s.u.).

Im **Frequenzbereich** sind zwei Methoden zur linearen Verzerrung üblich. Bei der ersten Variante werden die Koeffizienten eines Kurzzeitspektrums neu abgetastet und die Zwischenwerte durch lineare Interpolation gewonnen (siehe **Abbildung 7.4**). Die zweite Variante beruht auf der Voraussetzung, dass zur weiteren Verarbeitung eine Filterbank (z.B. nach der *Mel*- oder *Bark*-Skala) benutzt wird. Die meist nach einer (nicht-linearen) gehörsadäquaten Skala angeordneten Mittenfrequenzen und Bandbreiten der einzelnen Filter werden dann nochmals linear entsprechend dem Faktor verzerrt. Da man nach der Normierung eine konstante Anzahl von Koeffizienten erhalten möchte, ergibt sich für beide Verfahren das Problem, dass bei Verzerrungsfaktoren $\alpha_a > 1$ Werte zur Interpolation im oberen Frequenzbereich fehlen (siehe das Fragezeichen „?" in **Abbildung 7.4**). Eine Lösung dafür ist die Reduktion der Koeffizientenanzahl (Merkmale), die nach der Normierung verwendet werden. Dies würde allerdings alle Sprecher betreffen und bedeutet bei einer Sprachdatenbank mit gegebener Abtastrate einen erheblichen Verzicht auf Information, die noch in den Ausgangsdaten steckt. Eine andere Lösung geht daher einen Kompromiss ein und greift auf stückweise lineare Abbildungen zurück (s.u.). Der größte Teil des Spektrums bis $\beta \cdot (f_A/2)$ wird dabei nach wie vor entsprechend dem zugrundeliegenden Modell verzerrt.

Neben der Verzerrung im Zeit- und Frequenzbereich existieren auch noch andere Möglichkeiten, wie zum Beispiel für den **Cepstralbereich**. [Acero 1990] verwendet hierfür eine bilineare Transformation, die umgerechnet auf den Frequenzbereich folgende Form besitzt:

$$\omega^* = \omega + 2 \cdot \arctan\left(\frac{\gamma \cdot \sin(\omega)}{1 - \gamma \cdot \cos(\omega)}\right) \qquad \text{mit} \quad \omega = 2\pi \cdot \frac{f}{f_A} \tag{7.3}$$

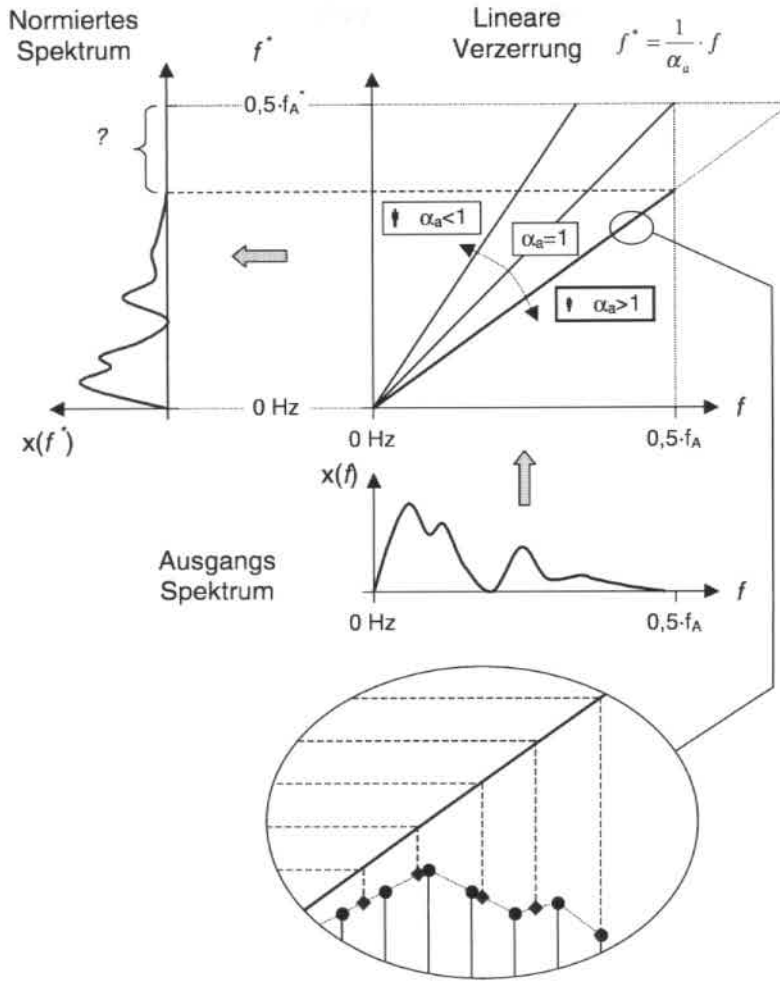


Abbildung 7.4: Verzerrung der Frequenzachse durch Neuabtastung mit linearer Interpolation.

Diese muss zunächst als eine Näherung für die Bark-Skala betrachtet werden (mit $\gamma \approx 0{,}6$). Acero verwendet dann einen Wertebereich für γ zwischen 0,4 und 0,6 um damit die Frequenzverzerrung durchzuführen. Ein Vorteil dieser Abbildung ist, dass sie das Ausgangsspektrum wieder in sich selbst transformiert, womit die oben angesprochenen Probleme entfallen. Es liegt allerdings für diese Abbildung kein Vokaltraktmodell zugrunde.

Für diese Arbeit wurden im dazu verwendeten JRTk (siehe Abschnitt 4.2) nur die Frequenzverfahren implementiert, da diese unabhängig von der weiteren Verarbeitung der Merkmale sind. Da auch gelegentlich unterschiedliche Filterbänke verwendet werden, bietet sich die Transformation direkt auf den Kurzzeit-Spektralkoeffizienten an.

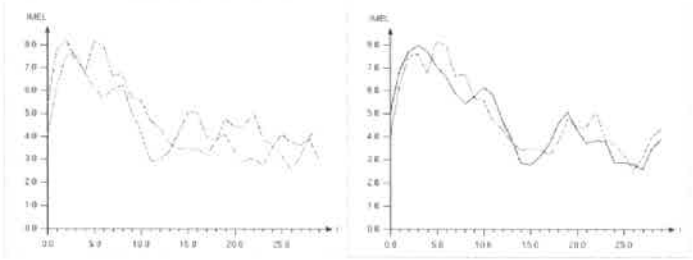


Abbildung 7.5: *Mel*-Filterbankkoeffizienten (logarithmische Darstellung, Index *i*) desselben Lautes zweier Sprecher mit unterschiedlichen Vokaltraktlängen. Links ohne Normierung, rechts mit Frequenztransformation eines Sprechers.

Abbildung 7.5 zeigt eine Darstellung der *Mel*-Filterbankkoeffizienten nach einer stückweise linearen Abbildung, die das Kurzzeit-Spektrum eines Sprechers auf das eines anderen transformiert (als typischer Wert für den Knickpunkt wurde hier $\beta = 0{,}8$ gesetzt, siehe **Abbildung 7.6**). Der Zusammenhang bei der **stückweise linearen Abbildung** (mit der Abtastrate f_A) ergibt sich zu:

$$f^* = \begin{cases} \frac{1}{\alpha_a} \cdot f & \text{für } f \le \beta \cdot \left(\frac{f_A}{2}\right) \\ \frac{\alpha_a - \beta}{\alpha_a - \alpha_a \beta} \cdot f + \frac{\beta - \alpha_a \beta}{\alpha_a - \alpha_a \beta} \cdot \left(\frac{f_A}{2}\right) & \text{für } f > \beta \cdot \left(\frac{f_A}{2}\right) \end{cases} \tag{7.4}$$

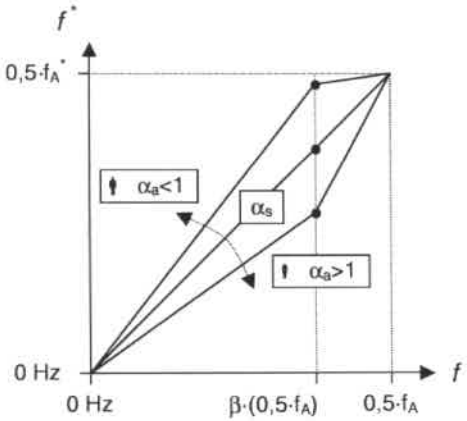


Abbildung 7.6: Frequenzverzerrung mit einer stückweise linearen Funktion.

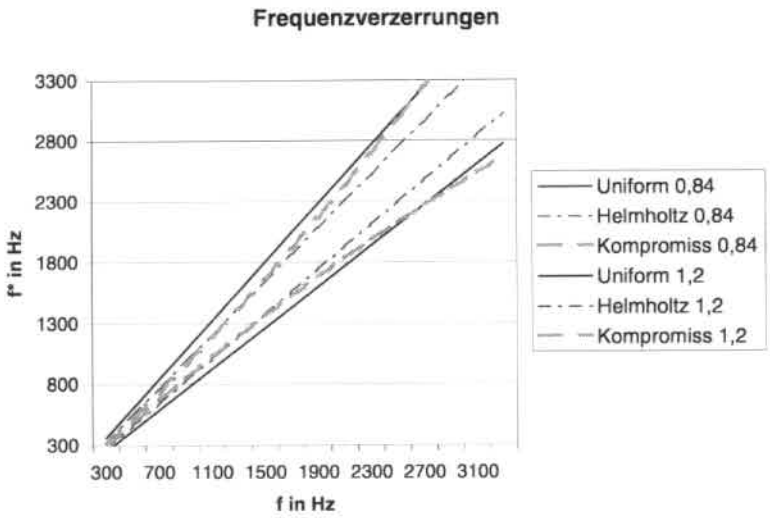


Abbildung 7.7: Lineare und nichtlineare Funktionen zur Frequenzverzerrung, jeweils für extreme Werte 0,84 und 1,2 des Verzerrungsfaktors: Röhrenmodell (Uniform), Helmholtz und eine Funktion aus [Eide und Gish 1996] (Kompromiss).

Ein großer Vorteil der Frequenzverfahren ist, dass neben dem linearen Modell auch leicht andere Modelle mit nichtlinearen Funktionen umgesetzt werden können. [Eide und Gish 1996] schlagen eine solche Funktion vor und normieren damit den dritten Formanten (s.u.: Formanten-VTLN):

$$f^* = \left(\frac{1}{\alpha_a}\right)^i \cdot f \qquad \text{mit} \quad i = \frac{3 \cdot f}{8000\,\text{Hz}} \tag{7.5}$$

Diese Funktion stellt einen Kompromiss zwischen dem einfachen Röhrenmodell (linear mit $i = 1$) und dem Helmholtz-Resonator (linear mit $i = 0{,}5$) dar. Dies gilt allerdings nur in einem eingeschränkten Frequenzbereich (siehe **Abbildung 7.7**), der in etwa der Telefonbandbreite entspricht. Auch hier müssen entsprechende Maßnahmen getroffen werden, damit die Zahl der normierten Merkmalskoeffizienten konstant bleibt. In [Zhan und Westphal 1997] ergaben sich für diese Funktion im Vergleich mit der stückweise linearen Abbildung keine besseren Ergebnisse.

Um mehr als eine Art der Frequenzverzerrung umzusetzen, wurde im Verlauf der hier beschriebenen Arbeit der Gedanke verfolgt, lautspezifisch die jeweils passende Normierungsfunktion zu verwenden. Für jeden Merkmalsvektor wurden dazu mehrere Versionen (eine pro Abbildung) erzeugt („Merkmalsströme", engl. *feature streams*). Bei der Auswertung der Merkmale mithilfe der akustischen Modelle des Erkenners wurden dann jeweils unterschiedliche Versionen verwendet und zwar je nach phonetischer Klassenzugehörigkeit des gerade betrachteten akustischen Modells. Hierfür wurden sowohl eine Kombination

verschiedener Abbildungen untersucht als auch nur die einfache lineare Abbildung in Kombination mit nicht verzerrten Merkmalen.

Die Zuordnung der unterschiedlichen Merkmalsströme zu den Lautgruppen erscheint vernünftig, da beispielsweise Merkmalsvektoren für „Pause" natürlich nicht vom Sprecher abhängen können. Genauso stellt sich die Frage, ob man stimmlose Laute verzerren soll, da für sie das Vokaltraktlängenmodell nicht gilt. Im Zusammenhang mit der Maximum-Likelihood-Bestimmung des Verzerrungsparameters (s.u.) zeigte sich die Verwendung der stückweise linear verzerrten Merkmale für alle Modelle aber überlegen. Lediglich mit speziellen unverzerrten Merkmalen für die Pausenmodelle ergaben sich leichte Vorteile. Daher wurden im Weiteren die Merkmale für alle Modelle stets mit der gleichen Verzerrungsfunktion berechnet.

7.3.3 Bestimmung des Verzerrungsfaktors

Auch für die Bestimmung des Verzerrungsfaktors gibt es diverse Möglichkeiten. Hier wird zunächst eine Übersicht verschiedener Ansätze gegeben, um dann im nächsten Abschnitt die Vorzüge und Nachteile der einzelnen Verfahren in Kombination mit den Verzerrungsfunktionen zu vergleichen:

Merkmalsbasierte Verfahren beruhen auf einem funktionalen Zusammenhang zwischen dem Verzerrungsfaktor und einem direkt aus dem Sprachsignal bestimmbaren Merkmal. Da man letztendlich an einer Normierung der Formanten F_i interessiert ist, liegt es nahe diese über eine längere Zeit zu mitteln und den so gewonnenen durchschnittlichen Formantwert $\overline{F_i}$ des Sprechers mit dem Wert $\overline{F_i}^*$ des Standardsprechers zu vergleichen. Mit dem linearen Verzerrungsmodell ergibt sich bei dieser **Formanten-VTLN** der Verzerrungsfaktor direkt aus dem Verhältnis der beiden Mittelwerte. Da es mehrere Formanten gibt (häufig werden $i = 1...3$ betrachtet), muss entweder einer dieser Werte selektiert oder mehrere Werte kombiniert werden. Die meisten Formanten hängen naturgemäß stark von den phonetischen Klassen ab, weshalb ein brauchbarer Mittelwert über relativ viele Beispiele geschätzt werden muss. F_3 ist von der Theorie her ein guter Kandidat, weil der Wert dieser Formante relativ konstant bleibt. Probleme bereitet vor allem die genaue Bestimmung der Formanten und der weiter oben schon erwähnte Umstand, dass das lineare Modell nicht für alle Laute gültig ist (besonders bei solchen ganz ohne Formanten). Daher spielt hier die ausgewählte Normierungsfunktion eine sehr entscheidende Rolle. Ein anderes Merkmal, das ebenfalls benutzt werden kann, ist die Grundfrequenz F_0. Anatomisch bedingt ergibt sich eine hohe Korrelation der Mittelwerte von Grundfrequenz (Stimmbänder) und Formanten (Vokaltrakt). Der Mittelwert von F_0 kann jedoch meist verlässlicher und mit weniger Daten bestimmt werden. Für die **F0-VTLN** muss allerdings eine geeignete Abbildungsfunktion auf den Verzerrungsfaktor gefunden werden.

Modellbasierte Verfahren beruhen auf einem akustischen Modell, also mit Dichteschätzungen (im Optimalfall für den Standardsprecher), und einer vorgegebenen Verzerrungsfunktion. Die Verzerrungsparameter werden so bestimmt, dass das normierte Signal möglichst gut zum Modell passt. Hierfür bietet sich das Maximum-Likelihood-Kriterium an (**ML-VTLN**). Nachteile sind der große Rechenaufwand und die Notwendigkeit eines akustischen Modells für Standardsprecher. Letzteres möchte man aber gerade erst durch

die Sprechernormierung erhalten. Für das Training mit der ML-VTLN bietet sich daher ein iteratives Vorgehen an. Ein Vorteil der ML-VTLN liegt darin, dass selbst bei unzureichendem Vokaltraktmodell (und sich daraus abgeleiteter Verzerrungsfunktion) die Verzerrungsparameter so gewählt werden, dass die Muster besser zum akustischen Modell passen. Daher ist die genaue Form der Normierungsfunktion weniger entscheidend als bei den merkmalsbasierten Verfahren. In dieser Arbeit wird noch ein weiterer modellbasierter Ansatz vorgestellt: die **LD-VTLN** (basierend auf der Linearen Diskriminanz), die vor allem die Nachteile der ML-VTLN im Training beseitigt.

7.3.4 Anwendung der VTLN

Bedingt durch die vielen Möglichkeiten für die Vokaltraktnormierung ist man auf experimentelle Vergleiche angewiesen. Über die letzten Jahre erfolgten einige Veröffentlichungen zu dieser Thematik:

[Andreou *et al.* 1994] stellten den ML-Ansatz zur Bestimmung des Verzerrungsfaktors α vor. Da die Umsetzung der ML-VTLN sehr rechenintensiv ist, wird in [Lee und Rose 1996] für einen Ziffernerkenner die Verzerrung mittels der *Mel*-Filterbank umgesetzt und statt des akustischen Modells des Erkenners mehrere Hilfsmodelle verwendet. Jede dieser einfachen Gaußmixturen ist mittels Trainingsdaten von Sprechern mit demselben Verzerrungsfaktor erstellt worden. Im Vergleich zur Bestimmung von α durch das Modell des Erkenners schneidet dieser Beschleunigungsversuch aber etwas schlechter ab. [Wegmann *et al.* 1996] verwenden eine stückweise lineare Verzerrung des Spektrums und zur beschleunigten Bestimmung des Faktors ebenfalls ein generisches akustisches Modell, das auf stimmhafte Sprachrahmen eingeschränkt wurde. Für die dort untersuchte Erkennung ganzer Dialoge werden im Mittel 52 Sekunden Sprachdaten für die ML-Bestimmung verwendet, was als wenig bezeichnet wird. Ein Vergleich mit einer Bestimmung über das Erkennermodell wird nicht gegeben.

Im Gegensatz dazu benutzen [Eide und Gish 1996] den dritten Formanten F_3, also eine merkmalsbasierte VTLN, mit einer linearen Verzerrung. Außerdem schlagen sie die bereits oben erwähnte nichtlineare Verzerrungsfunktion vor, die bei ihren Experimenten gegenüber der linearen Normierung bessere Ergebnisse liefert. Für die Bestimmung des Medianwertes von F_3 eines Sprechers verwenden sie über 60 Sekunden Sprachmaterial. [Lincoln *et al.* 1997] benutzen die ersten beiden Formanten zur Sprechernormierung bei der Phonemerkennung, bestimmen den Verzerrungsfaktor aber über einen ML-Ansatz. Die zwei ersten Formanten einzelner Vokale werden dazu durch zwei eindimensionaler Gaußdichten beschrieben. Sie tragen zwar damit der unzuverlässigen Bestimmung der Formanten Rechnung, benötigen aber für alle Phonemklassen Hilfsmodelle und eine erste Zuordnung der stimmhaften Muster zu den entsprechenden Vokalklassen.

[Gouvea und Stern 1997] verwenden alle drei Formanten um eine lineare Abbildungsfunktion zu finden und vergleichen diesen Ansatz mit einer ML-Schätzung nach [Wegmann *et al.* 1996]. Sie finden zwar heraus, dass in ihrem Fall die Formanten-VTLN deutlich besser ist, geben aber als Richtlinie etwa fünf Äußerung an, um die Medianschätzung der Formanten zu stabilisieren. Bei nur einem Satz bricht die Erkennungsrate durch falsche Normierung abrupt ein und ist schlechter als ohne Normierung. In [Zhan und Westphal 1997] werden ebenfalls

die beiden Ansätze verglichen und auch Ergebnisse für unterschiedliche Verzerrungsfunktionen (stückweise linear und nichtlinear) gegeben. Die Ergebnisse, bei denen äußerungsweise ein Normierungsfaktor bestimmt wurde, zeigen durchweg einen Vorteil mit der linearen Verzerrung und ergeben eine Überlegenheit der ML-VTLN gegenüber der Formanten-VTLN mittels einer der ersten drei Formanten.

In [Haiber 1998] wird statt der Formanten als Merkmal die mittlere Grundfrequenz F_0 des Sprechers bestimmt. Der Verzerrungsfaktor liegt damit zunächst noch nicht fest, sondern wird beispielsweise über eine Regression mit dem optimalen Verzerrungsfaktor ermittelt, wofür dort einige Möglichkeiten aufgezeigt werden. Im Gegensatz zu den bisher betrachteten Verfahren wird dieser optimale Verzerrungsfaktor direkt über die Wortfehlerrate auf einem Kreuzvalidierungs-Testset bestimmt (prinzipiell könnte man dieses Vorgehen auch bei der Formanten-VTLN anwenden). Die Bestimmung der Grundfrequenz kann robuster als die der Formanten erfolgen und ist im Vergleich mit der aufwändigen ML-Bestimmung wesentlich schneller. Aus diesem Grund wird diese alternative Möglichkeit in [Haiber 1998] entwickelt, obwohl sie etwas schlechtere Ergebnisse als die ML-VTLN liefert. Ein Problem bei dieser Methode liegt in den Verzerrungsfaktoren für das Training. Die weiter unten vorgestellte LD-VTLN würde dieses Problem lösen. In [Schubert 1999] wurden als Referenzwerte für die Regression die Verzerrungsfaktoren eines ML-VTLN-Trainings verwendet. Dort wurde mit der F0-VTLN ebenfalls die Erkennung beschleunigt, die Ergebnisse der ML-VTLN jedoch nicht erreicht.

Beim Vergleich der Methoden kommt es sicherlich auf viele Randbedingungen an. Aus eigenen Versuchen und den Ergebnissen anderer Gruppen kann man jedoch folgende groben Auswertung ziehen: Für die Formanten-VTLN und F0-VTLN liegen die Verbesserungen bei der Erkennung spontaner Sprache bei etwa 5 % Fehlerreduktion [Eide und Gish 1996, Zhan und Westphal 1997, Haiber 1998, Schubert 1999], bei der ML-VTLN meist etwa um 10 % [Wegmann *et al.* 1996, Finke *et al.* 1997, Zhan und Westphal 1997]. Die ML-VTLN ist somit der erfolgreichere Ansatz, weist aber auch den Nachteil einer aufwändigen Berechnung auf. Die Schätzung des Verzerrungsfaktors erfolgt meist über eine relativ lange Dauer von Testdaten, kann aber durch Einschränkung auf stimmhafte Regionen robuster gemacht werden.

7.4 ML-VTLN

Aufgrund des Erfolges der ML-VTLN bei der Sprechernormierung wurde dieser Ansatz auch für die hier vorgestellten Untersuchungen verwendet. Die speziellen Anforderungen, die dabei erfüllt werden müssen, nämlich schnelle und äußerungsweise Normierung, wurden bereits auf ähnlichen Anwendungen vom Autor entwickelt und erprobt:

- Eine Beschleunigung des Verfahrens wird dadurch erreicht, dass die Zuordnung der Analyserahmen und der HMM-Zustände nur einmal erfolgt.
- Eine robustere Schätzung des Verzerrungsfaktors ergibt sich aus der Beschränkung der Likelihood-Auswertung auf stimmhafte Zustände.

Schauen wir uns dazu den Ablauf der Erkennung und das Training mit der ML-VTLN an:

Erkennung

Erkennung mit ML-VTLN

1. Da noch kein Verzerrungsfaktor α für den Sprecher bekannt ist, wird dieser zunächst auf einen mittleren Wert (z.B. 1,0) gesetzt oder durch ein merkmalsbasiertes Verfahren grob vorbestimmt. Eine erste Hypothese w (Wortfolge) wird aufgrund der so berechneten Merkmalsvektorfolge durch den Erkenner generiert.
2. Mit der Hypothese w wird eine Viterbi-Zuordnung z gefunden, um die Likelihood für unterschiedlich verzerrte Merkmalsvektorfolgen $\mathbf{X}_\alpha$ berechnen und vergleichen zu können. Der optimale Verzerrungsfaktor α^* wird über den maximalen Likelihood-Wert (siehe Abschnitt 3.3.4) bestimmt:

$$\alpha^* = \arg\max_{\alpha}(p(\mathbf{X}_\alpha \mid z, \lambda)) \tag{7.6}$$

3. Der nachfolgende Erkennungslauf mit den durch α^* sprechernormierten Merkmalen führt zur finalen Hypothese w^*.

Bei der Likelihood-Bestimmung in Schritt 2 wird auf die Umrechnung des Modells λ wie in Gleichung (3.30) mithilfe der Jacobischen Determinante (siehe Seite 91) verzichtet, da diese von α abhängt und der Zusammenhang im betrachteten Merkmalsraum sehr komplex ist. Dennoch lassen sich so in der Praxis gute Ergebnisse erzielen, weshalb dieses Vorgehen gebräuchlich ist.

Der Wertebereich für α, der im Schritt 2 untersucht wird, ergibt sich aus dem iterativen Training der VTLN-Verzerrungsfaktoren, das weiter unten beschrieben wird. Ein typisches Set umfasst 13 Werte um den Faktor 1,0 in gleichmäßigen Abständen von 0,02. Für den Schritt 2 der Erkennung müssten dann idealerweise 13 Viterbi-Zuordnungen mit der Hypothese w durchgeführt werden. Es reicht aber als Näherung diese Zuordnung nur einmal durchzuführen (mit den Merkmalen, die auch in Schritt 1 benutzt wurden) und dann lediglich die Merkmalsvektorfolgen $\mathbf{X}_\alpha$ neu zu bewerten. Es werden hierfür, wie oben schon angedeutet, nur die Zustände (und damit nur die zugeordneten Merkmalsvektoren) von stimmhaften Lauten verwendet. Die für alle α übereinstimmende Zuordnung hat auch den Vorteil, dass die (auf stimmhafte Laute eingeschränkte) Anzahl der ausgewerteten Merkmalsvektoren immer gleich ist. So lassen sich die berechneten Likelihood-Werte besser vergleichen.

Training

Für das Training wurde ein ähnliches Vorgehen wie bei der Erkennung gewählt. Zunächst sind keine Verzerrungsfaktoren bekannt, also wird mit 1,0 begonnen. Nach einer Viterbi-Zuordnung mit diesem Ausgangswert wird ein neuer Verzerrungsfaktor entsprechend der maximalen Likelihood bestimmt. Diesmal ergeben sich jedoch zwei Abweichungen:

1. Da der Verzerrungsfaktor als konstant für den Sprecher angenommen wird, kann seine Bestimmung über alle Äußerungen des Trainingssprechers erfolgen.

2. Nachdem für alle Sprecher ein Verzerrungsfaktor gefunden wurde, wird mit den entsprechend verzerrten Trainingsmustern das akustische Modell des Erkenners neu berechnet. Mit dem neuen Modell kann wieder eine Likelihood-Bewertung vorgenommen werden, um die Faktoren für jeden Sprecher abermals erneut zu bestimmen.

So ergeben sich über die Iterationen abwechselnd Verzerrungsfaktoren und akustische Modelle, die idealerweise mit immer besser normierten Merkmalen bestimmt werden. **Abbildung 7.8** zeigt die Entwicklung der Verzerrungsfaktoren über die Trainingsiterationen. Zunächst wurde der Erkenner über 4 EM-Iterationen ohne VTLN trainiert (siehe Standardtraining im Abschnitt 4.2, S. 66) und dann über mehrere Iterationen (5 bis 11) mit der ML-VTLN. Die Verzerrungsfaktoren der männlichen und weiblichen Sprecher gehen dabei immer weiter auseinander und es ergibt sich insgesamt die erwartete bimodale Verteilung.

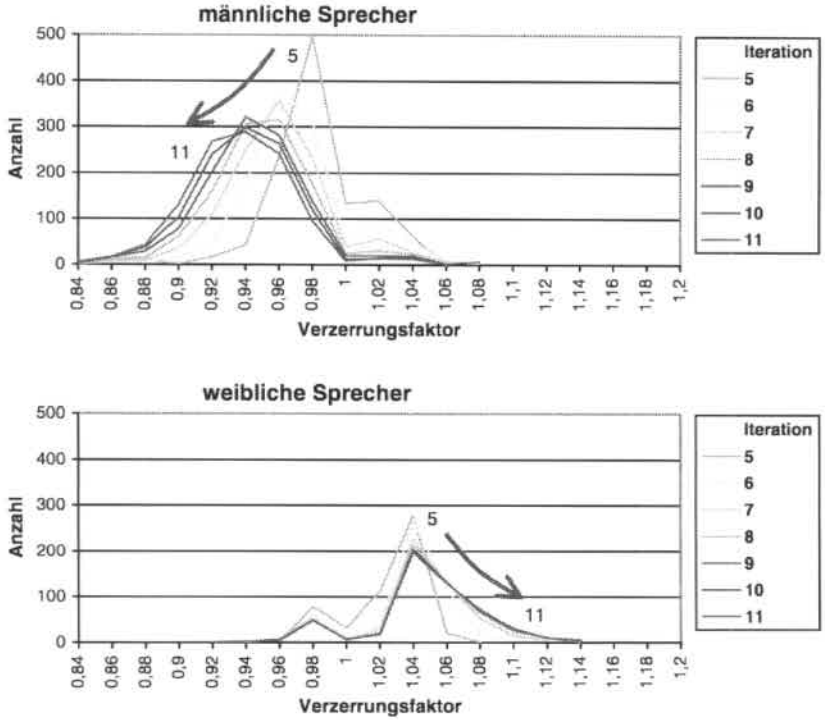


Abbildung 7.8: Histogramme (Sprecheranzahl) der Verzerrungsfaktoren über die Iterationen für die ML-VTLN.

7.5 LD-VTLN

Nachteile der ML-VTLN

Trotz des Erfolges der ML-VTLN und der besseren Ergebnisse gegenüber den merkmalsbasierten Verfahren hat die ML-VTLN gewisse Nachteile. Neben der bereits erwähnten Verarbeitungsgeschwindigkeit finden sich weitere Schwächen bedingt durch das iterative Training:

1. Die Verzerrungsfaktoren erleiden zuweilen eine Drift von Iteration zu Iteration (die auch [Acero 1990] und [Lee und Rose 1996] erwähnen). Dadurch, dass sich Modell und Verzerrungsfaktor über die Iterationen gegenseitig bedingen, kann der gesamte Merkmalsraum „zusammenfallen". Dies hat seine Ursachen in der Nichtbeachtung der Jacobischen Matrix bei der Maximum-Likelihood-Bestimmung der Verzerrungsfaktoren. Über die oben gewählte Näherung kann sich ein globales Optimum beispielsweise so ergeben, dass alle Muster und dadurch die Mittelwerte der Dichten auf einen konstanten Vektor abgebildet werden. Dies führt zwar zu einer hohen Likelihood-Bewertung nach Gleichung (7.6), ist aber nicht Sinn der Normierung.

2. Ebenso können lediglich kleinere, lokale Maxima erreicht werden, die zwar eventuell das Ergebnis verbessern, aber nur einen Teil des möglichen Gewinns bringen. Zu Beginn müssen ja zumindest zwei große Sprechergruppen, nämlich Männer und Frauen, innerhalb des noch nicht normierten akustischen Modells repräsentiert sein. Nun könnte es passieren, dass sich für die meisten Männer ein Verzerrungsfaktor ergibt, der ihre Sprachmerkmale auf die eines durchschnittlichen männlichen Sprechers normiert. Geschieht dies auch entsprechend für die weiblichen Sprecher, bilden sich zwei Gruppen, die zwar jeweils innerhalb normiert sind, aber eigene Parameter für die akustische Modellierung benötigen. Die Verzerrungsfaktoren wären für dieses Beispiel nicht mehr proportional zur Vokaltraktlänge, sondern nur zur Stimmlage innerhalb der Geschlechter. Als dies in [Wegmann *et al.* 1996] absichtlich und mithilfe getrennter Modelle versucht wurde, ergaben sich zwar etwas bessere Erkennungsraten als für ein nicht normiertes System, aber schlechtere als für einen insgesamt sprechernormierten Erkenner.

3. Ein Nachteil betrifft Systeme, die eine LDA benutzten (siehe Abschnitt 5.4.2). Die Transformationsmatrix zur Dimensionsreduktion und Normierung der Merkmale wird dabei für bestimmte Klassen und gegebene Sprachmuster bestimmt. Durch die sich ändernden Verzerrungsfaktoren ändern sich über die Trainingsiterationen nun aber auch die Muster. Da gewisse sprecherbedingte Varianz aus den Merkmalen herausgenommen wird, ist die verwendete Transformation nicht mehr optimal nach dem LDA-Kriterium. Wird sie neu bestimmt, erfordert dies auch eine neue Initialisierung des akustischen Modells mit anschließendem Training.

Um auszuschließen, dass sich (wie unter 2. erwähnt) mehrere Sprechergruppen bilden und um gleichzeitig die Berechnungsgeschwindigkeit zu erhöhen, kann man statt des Erkennermodells ein Hilfsmodell λ_l mit einzelnen Gaußdichten für die phonetischen Klassen verwenden [Westphal *et al.* 1998, Welling *et al.* 1999]. Werden allerdings wie oben die

Verzerrungsfaktoren und das akustische Modell abwechselnd bestimmt, verstärkt sich das Driftproblem, was nun mit einem vereinfachenden Beispiel nochmals genauer erläutert wird. Diese Überlegung diente als Motivation für die anschließend vorgestellte LD-VTLN:

Angenommen die Klassenzuordnungen sind festgelegt und die Modelle dafür besitzen einzelne Gaußdichten mit einer festen Kovarianzmatrix. Letztere wurde zur Einheitsmatrix festgelegt, während die Mittelwerte sich durch die Trainingsmuster der jeweiligen Klasse bestimmen. Dann entspricht die Log-Likelihood der Muster mit dem akustischen Modell λ_1 (bis auf Vorzeichen und Konstanten) der Spur der mittleren Klassen-Kovarianzmatrix **W** (siehe Abschnitt 3.3.3). Durch den Vorzeichenunterschied ergibt sich aus der Minimierung der Spur von **W** eine Maximierung der Likelihood. Werden durch iterative Bestimmung der Sprechernormierung alle Trainingsmuster so verzerrt, dass sie aufeinander fallen, gilt das ebenso für die Mittelwertvektoren der Modelle. Dadurch verschwinden natürlich die Varianzen der Klassen und das Spur-Kriterium erreicht seinen Minimalwert Null. Dennoch stellt dies nicht das gewünschten Ergebnis dar. Da die Klassen unabhängig voneinander betrachtet werden, werden zwar die Muster innerhalb einer Klasse ähnlicher, lassen sich aber nicht mehr von denen der anderen unterscheiden.

Ein neues Kriterium für die VTLN

Daher wird nun mittels eines für diese Anwendung neuen Kriteriums die iterative Bestimmung der Verzerrungsfaktoren verbessert. Verwendet man die in Abschnitt 3.3.3 vorgestellte Lineare Diskriminanz *LD*, verhindert die zusätzliche Kovarianzmatrix aller Merkmalsvektoren **T** ein Zusammenfallen der Merkmale durch die iterative Bestimmung der Verzerrungsfaktoren:

$$LD = \frac{|\mathbf{T}|}{|\mathbf{W}|} \tag{7.7}$$

Diese Art der Sprechernormierung wurde für die vorliegende Arbeit neu entwickelt und erstmals in [Westphal *et al.* 1998] vorgestellt. Wie bei der Linearen Diskriminanzanalyse möchte man auch durch die Sprechernormierung erreichen, dass die Muster einer Klasse näher zusammen rücken und die Muster insgesamt noch eine hohe Varianz (also Unterscheidbarkeit) aufweisen. Der Unterschied ist aber:

- Bei der LDA wird der Merkmalsraum als Ganzes *linear* transformiert, so dass die ersten Koeffizienten einen großen Wert für *LD* ergeben und so die Dimension reduziert werden kann.

- Bei der LD-VTLN werden die Muster einzelner Sprecher durch die Normierung *nichtlinear* verändert um den Wert für *LD* zu maximieren.

Eine Bemerkung dazu: Die Aussage über die nichtlineare Veränderung bezieht sich nicht auf den Spektralbereich (für den die Sprechernormierung in der Regel sogar linear ist), sondern auf die cepstralen Merkmale auf denen hier der *LD*-Wert bestimmt wird. Die nichtlineare Abbildung der cepstralen Merkmale durch die VTLN ist ein Grund, warum diese auch in Kombination mit der Sprechernormierung durch MLLR die Erkennungsrate weiter verbessern kann (siehe dafür z.B. [Zhan *et al.* 1997]).

Wie ergeben sich nun aber beim Training mit der LD-VTLN die Verzerrungsfaktoren α_a der einzelnen Sprecher a? Die entsprechend verzerrten Sprachmuster eines Sprechers gehen bei der Berechnung der beiden Matrizen **W** und **T** ein, aber ebenso die aller anderen Sprecher. Da der Zusammenhang der finalen Merkmale mit dem Verzerrungsfaktor α_a sehr komplex ist, kann man im Allgemeinen keine geschlossene Lösung angegeben. Daher wird nun ein iteratives Verfahren beschrieben, mit dem sich ähnlich wie bei der ML-VTLN für jeden Sprecher ein optimaler Verzerrungsfaktor finden lässt.

Da die Berechnung der beiden Kovarianzmatrizen **W** und **T** alle Sprachmuster der Trainingsmenge verlangt, erfolgt die Berechnung aus praktischen Überlegungen indirekt über

die Klassenakkumulatoren $\mathbf{m}_s = \sum_{k\,|\,z[k]=s} \mathbf{x}[k]$ für alle Klassen $s = 1 \ldots S$, (7.8)

die Akkumulatormatrix $\mathbf{S} = \sum_k \mathbf{x}[k] \cdot \mathbf{x}[k]^T$ (7.9)

und die Anzahl n_s der Merkmalsvektoren der jeweiligen Klasse.

Daraus können jederzeit die Matrizen **W** und **T** berechnet werden:

$$n = \sum_{s=1}^{S} n_s\,, \quad \mathbf{m} = \sum_{s=1}^{S} \mathbf{m}_s \tag{7.10}$$

mit $\boldsymbol{\mu} = \frac{1}{n}\sum_{k=1}^{n} \mathbf{x}[k] = \frac{1}{n} \cdot \mathbf{m}$ und $\boldsymbol{\mu}_s = \frac{1}{n_s} \sum_{k\,|\,z[k]=s} \mathbf{x}[k] = \frac{1}{n_s} \cdot \mathbf{m}_s$ erhält man:

$$\begin{aligned}
\mathbf{T} &= \frac{1}{n}\sum_{k=1}^{n} (\mathbf{x}[k]-\boldsymbol{\mu}) \cdot (\mathbf{x}[k]-\boldsymbol{\mu})^T = \frac{1}{n}\sum_{k=1}^{n} \left(\mathbf{x}[k] \cdot \mathbf{x}[k]^T - \boldsymbol{\mu} \cdot \mathbf{x}[k]^T - \mathbf{x}[k] \cdot \boldsymbol{\mu}^T + \boldsymbol{\mu} \cdot \boldsymbol{\mu}^T\right) \\
&= \frac{1}{n}\sum_{k=1}^{n} \left(\mathbf{x}[k] \cdot \mathbf{x}[k]^T\right) - \boldsymbol{\mu} \cdot \boldsymbol{\mu}^T - \boldsymbol{\mu} \cdot \boldsymbol{\mu}^T + \boldsymbol{\mu} \cdot \boldsymbol{\mu}^T = \frac{1}{n}\mathbf{S} - \boldsymbol{\mu} \cdot \boldsymbol{\mu}^T \\
&= \frac{1}{n}\left(\mathbf{S} - \frac{1}{n}\mathbf{m} \cdot \mathbf{m}^T\right)
\end{aligned} \tag{7.11}$$

$$\begin{aligned}
\mathbf{W} &= \sum_{s=1}^{S} \frac{n_s}{n} \mathbf{W}_s = \sum_{s=1}^{S} \frac{n_s}{n} \frac{1}{n_s} \sum_{k\,|\,z[k]=s} (\mathbf{x}[k]-\boldsymbol{\mu}_s) \cdot (\mathbf{x}[k]-\boldsymbol{\mu}_s)^T \\
&= \sum_{s=1}^{S} \frac{n_s}{n} \left(\frac{1}{n_s} \sum_{k\,|\,z[k]=s} \left(\mathbf{x}[k] \cdot \mathbf{x}[k]^T\right) - \boldsymbol{\mu}_s \cdot \boldsymbol{\mu}_s^T\right) = \frac{1}{n}\sum_{s=1}^{S} \sum_{k\,|\,z[k]=s} \left(\mathbf{x}[k] \cdot \mathbf{x}[k]^T\right) - \sum_{s=1}^{S} \frac{n_s}{n} \left(\frac{\mathbf{m}_s}{n_s} \cdot \frac{\mathbf{m}_s^T}{n_s}\right) \\
&= \frac{1}{n}\left(\mathbf{S} - \sum_{s=1}^{S} \frac{1}{n_s} \mathbf{m}_s \cdot \mathbf{m}_s^T\right)
\end{aligned} \tag{7.12}$$

Der Vorteil ergibt sich nun daraus, dass der Beitrag eines einzelnen Sprechers mit dem Verzerrungsfaktor α_a leicht entfernt werden kann. Dazu muss nur jeder seiner mit α_a verzerrten Merkmalsvektoren vom entsprechenden Klassenakkumulator und der

Akkumulatormatrix abgezogen werden. Umgekehrt können die Merkmalsvektoren, die durch einen anderen Verzerrungsfaktor erzeugt wurden, wieder zu den Akkumulatoren hinzugefügt werden.

Das Schema einer Verzerrungsfaktoriteration für alle Sprecher wird nun beschrieben. Die Viterbi-Zuordnungen $z = z[1]...z[K]$ wurden wie beim Standardtraining (Abschnitt 4.2) festgehalten, könnten aber auch für jede Iteration neu bestimmt werden. Mit der hier verfolgten Methode sind die Verzerrungsfaktoren jedoch lediglich von der Zuordnung und den verwendeten Merkmalen abhängig, weshalb das akustische Modell eines Erkenners nicht gebraucht wird.

Eine Trainingsiteration mit LD-VTLN:

1. Berechne die Akkumulatoren A für alle Sprachmuster mit den aktuellen Verzerrungsfaktoren der jeweiligen Sprecher.

2. Für jeden Sprecher:

- 2.1: Verzerre die Sprachmuster des Sprechers mit seinem aktuellen Verzerrungsfaktor und entferne seinen Beitrag von den Akkumulatoren A aus Schritt 1. Daraus entstehen die Akkumulatoren A'.
- 2.2: Verzerre die Merkmale für jeden zu untersuchenden Verzerrungsfaktor α, füge sie zu A', berechne die Matrizen **T** und **W** für diesen Fall und daraus die Lineare Diskriminanz *LD*.
- 2.3: Ersetze den aktuellen Verzerrungsfaktor des Sprechers mit demjenigen α, das den größten *LD*-Wert erreichte.

Werden die Akkumulatoren A nach Schritt 2.3 mit dem neu geschätzten Verzerrungsfaktors des Sprechers angepasst (was leicht mithilfe von A' möglich ist), ist garantiert, dass der *LD*-Wert monoton wächst bis ein lokales Optimum erreicht wird. Auch mit dem obigen Schema, das die Akkumulatoren erst nach einem Durchlauf durch die Trainingsdaten neu bestimmt, konnte in der Praxis ein schnelles Konvergenzverhalten erzielt werden.

Abbildung 7.9 zeigt (wie schon oben für die ML-VTLN) ein Histogramm der mit der LD-VTLN gefundenen Verzerrungsfaktoren. Diese nehmen im Vergleich mit **Abbildung 7.8** bereits von Anfang an weiter auseinanderliegende Werte an. Diesmal verändern sie sich nach der zweiten VTLN-Iteration kaum noch. Auch die Differenzen zu den ersten, von 1,0 verschiedenen Verzerrungsfaktoren sind nach einigen Iterationen wesentlich geringer als mit der ML-VTLN (**Abbildung 7.10**).

Das Trainingsverhalten der LD-VTLN ist also wesentlich vorteilhafter als das der ML-VTLN. Die Verzerrungsfaktoren bleiben stabil, werden schneller gefunden und sind nicht abhängig vom akustischen Modell eines Erkenners. Die Merkmale (zur Berechnung der Linearen Diskriminanz) können daher frei gewählt werden, sollten sich aber an diejenigen anlehnen, die auch später zur Modellierung im Erkenner benutzt werden. Um die Abhängigkeit von einer speziellen LDA-Transformationsmatrix zu umgehen (siehe Nachteil 3 der ML-VTLN), können deren Ausgangsmerkmale (also der erweiterte Merkmalsvektor nach Gleichung (2.14), Seite 22) verwendet werden.

Mit größerem Aufwand kann auch die Lineare Diskriminanz der dimensionsreduzierten Merkmale (Zielraum der LDA) optimiert werden. Die jeweils optimale Transformationsmatrix bei der Bestimmung des Verzerrungsfaktors eines Sprechers ergibt sich fast als Nebenprodukt der LD-VTLN, denn dafür ist lediglich noch die simultane Diagonalisierung der bereits vorliegenden Matrizen **W** und **T** nötig (vergleiche Abschnitt 5.4.2). Die Bestimmung der Eigenwerte und Eigenvektoren ist allerdings relativ rechenaufwändig und die Optimierung im Zielraum hängt von der Anzahl der dort verwendeten Merkmale ab. Daher wurde diese Möglichkeit nicht verwendet, sondern die LD-VTLN mit dem erweiterten Merkmalsvektor durchgeführt.

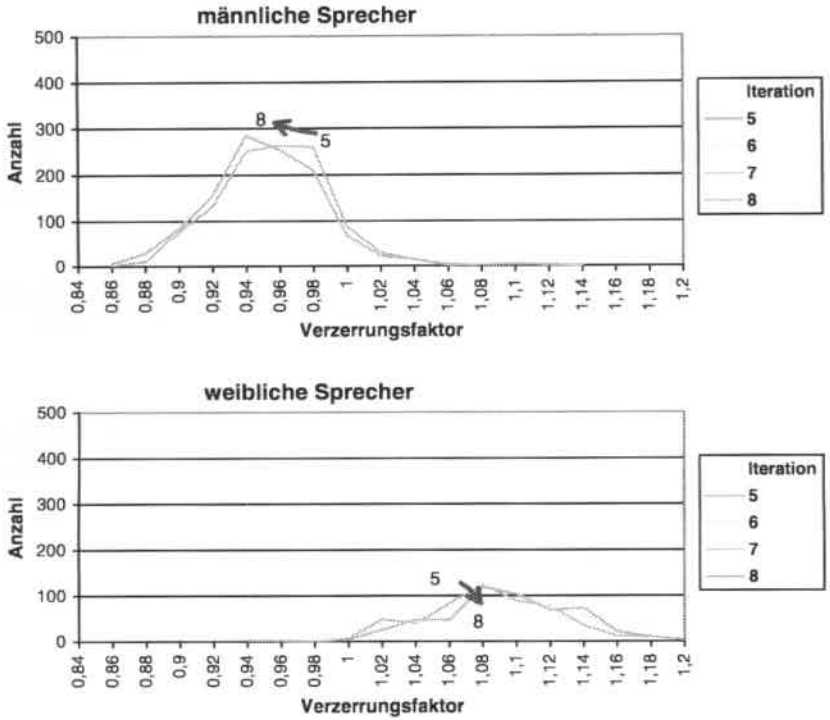


Abbildung 7.9: Histogramme (Sprecheranzahl) der Verzerrungsfaktoren über die Iterationen für die LD-VTLN.

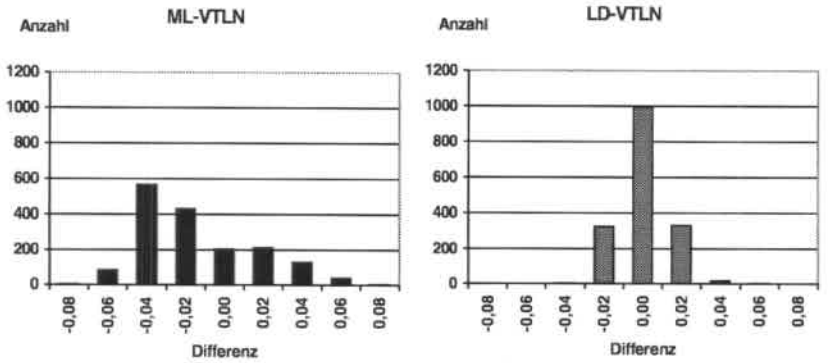


Abbildung 7.10: Histogramme (Sprecheranzahl) der Differenz des Verzerrungsfaktors zwischen der ersten und vierter VTLN-Iteration (entspricht den Iterationen Nummer 5 und 8).

7.6 Versuchsergebnisse

Für Experimente mit der Sprechernormierung wurden zunächst entsprechende Erkenner aufbauend auf dem Basiserkenner erzeugt. Über zusätzliche Iterationen mit der ML-VTLN und der LD-VTLN wurden zunächst Verzerrungsfaktoren V_1 bestimmt. Danach erfolgte ein Standardtraining inklusive LDA um neue Erkenner zu trainieren. Da bei der LD-VTLN die Verzerrungsfaktoren nicht vom Erkenner und der speziellen LDA-Matrix abhängen, wurde V_1 mit drei LD-VTLN-Iterationen berechnet. Für die ML-VTLN wurden zunächst vier Iterationen für V_1 verwendet und nach der Neuberechnung der LDA bei allen vier Trainingsiterationen die Verzerrungsfaktoren nochmals neu bestimmt.

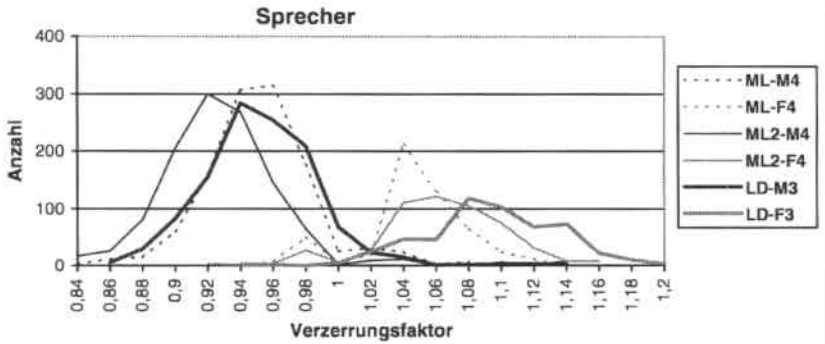


Abbildung 7.11: Vergleich der Verzerrungsfaktor-Histogramme.

ML: Nach 4 Iterationen mit ML-VTLN.

ML2: Nach neuer Initialisierung des akustischen Modells und 4 weiteren Iterationen mit ML-VTLN.

LD: Nach 3 Iterationen mit LD-VTLN.

M*i* und F*i*: Männliche und weibliche Sprecher nach *i* Iterationen.

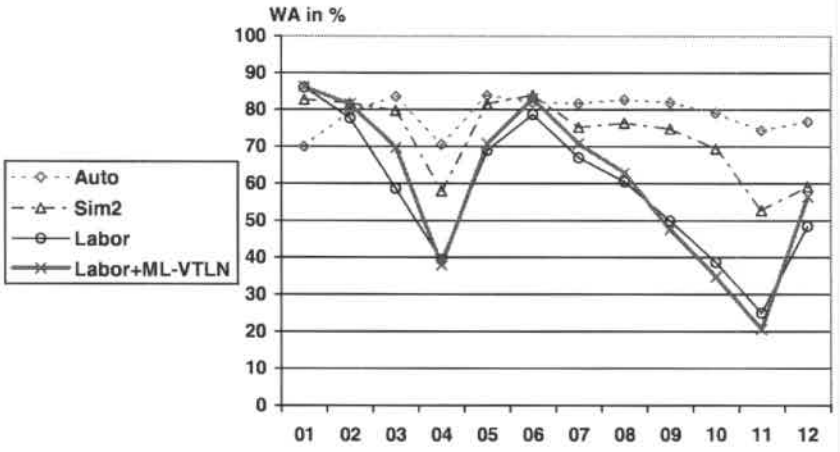
Abbildung 7.11 zeigt einen Vergleich der so erhaltenen Verteilungen der Verzerrungsfaktoren über die Sprecher. Erst nachdem bei der ML-VTLN die Verzerrungsfaktoren ein zweites Mal iterativ bestimmt wurden, ergibt sich eine ähnlich weit gestreckte Verteilung wie bei der LD-VTLN. Die absolute Lage ist jedoch um einiges hin zu kleineren Werte verschoben.

Tabelle 7.1 zeigt die Ergebnisse für das Testset der rauscharmen Kategorie 01 mit dem Trainingsmikrofon (Sennheiser). Während im Training die Verzerrungsfaktoren sprecherweise bestimmt werden konnten, wurden bei der Erkennung nur die Muster der aktuellen Äußerung zur Verfügung gestellt. Um auch auf diesen wenigen Daten gute Verzerrungsfaktoren zu erhalten, benutzt die hierzu verwendete ML-VTLN nur stimmhafte Merkmalsvektoren (s.o.). Für beide sprechernormierten Systeme ergeben sich gegenüber dem Basiserkenner ohne VTLN deutliche Fehlerreduktionen. Die LD-VTLN liegt dabei mit 19 % vor der ML-VTLN mit 13 %.

Tabelle 7.1: Ein Vergleich zwischen ML und LD Kriterium zur Bestimmung des Verzerrungsfaktors für die VTLN im Training.

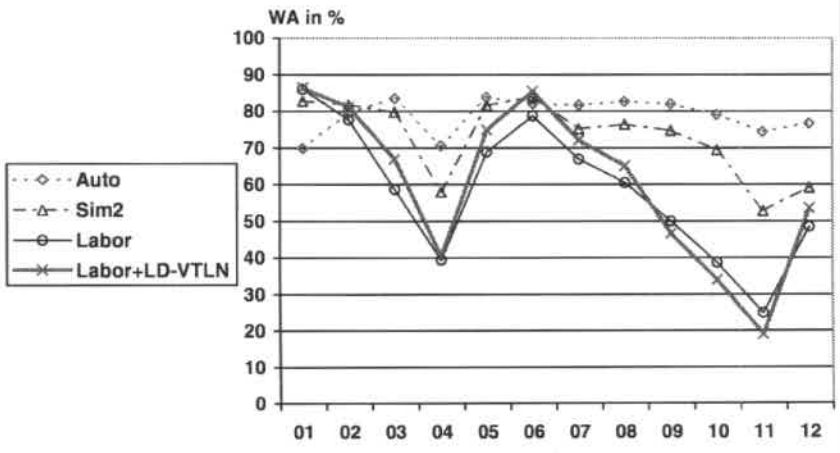
Sprechernormierung:		WA in %	relative Fehlerreduktion
Training sprecherweise	Test äußerungsweise		
keine	keine	86,9	-
ML-VTLN	ML-VTLN	88,6	13 %
LD-VTLN	ML-VTLN	89,4	19 %

Die Ergebnisse für das im Fahrzeug installierte Mikrofon für die verschiedenen Kategorien zeigen **Abbildung 7.12** für die ML-VTLN und **Abbildung 7.13** für die LD-VTLN. Hierbei ist ebenfalls die LD-VTLN überlegen und erreicht Fehlerreduktionen bis über 30 %. Allerdings wird die Bestimmung des Verzerrungsfaktors für stark gestörte Kategorien (vor allem 09 bis 11) immer unzuverlässiger und die Ergebnisse dadurch schlechter als mit dem Basiserkenner. Das liegt zum einen an den schlechteren Ausgangshypothesen (häufigere Falschzuordnung), zum anderen werden gestörte Merkmalvektoren mit dem akustischen Modell für ungestörte Sprache bewertet. Wird so ein falscher Verzerrungsfaktor gefunden, stimmen die damit verzerrten Merkmale noch weniger mit dem sprechernormierten Modell der jeweils richtigen Klasse überein. Dennoch kann man davon ausgehen, dass im Allgemeinen richtig normierte Merkmale robuster gegenüber additiven Störungen sind, da die Muster verschiedener Klassen im Schnitt eine größere Distanz haben. Das ist wohl der Grund dafür, warum für einige leicht gestörte Kategorien relativ große Fehlerreduktionen erreicht wurden.



Kategorie / System	01 aus	02 Blinker	03 Fenster	04 Lstark	05 Lmittel	06 0km/h	07 25km/h	08 50km/h	09 75km/h	10 100km/h	11 125km/h	12 beschl.
Basiserkenner (WA in %)	86,1	77,7	58,6	39,4	68,9	78,7	66,9	60,5	50,0	38,7	25,0	48,5
mit ML-VTLN (WA in %)	86,2	81,5	69,8	38,0	70,8	83,0	70,7	62,9	47,6	34,8	20,6	56,4
ER relativ	1%	17%	27%	-2%	6%	20%	11%	6%	-5%	-6%	-6%	15%

Abbildung 7.12: Ergebnisse der ML-VTLN.



Kategorie / System	01 aus	02 Blinker	03 Fenster	04 Lstark	05 Lmittel	06 0km/h	07 25km/h	08 50km/h	09 75km/h	10 100km/h	11 125km/h	12 beschl.
Basiserkenner (WA in %)	86,1	77,7	58,6	39,4	68,9	78,7	66,9	60,5	50,0	38,7	25,0	48,5
mit LD-VTLN (WA in %)	86,4	81,2	66,9	40,5	74,9	85,5	72,1	65,1	46,6	34,0	19,1	53,6
ER relativ	2%	16%	20%	2%	19%	32%	16%	12%	-7%	-8%	-8%	10%

Abbildung 7.13: Ergebnisse der LD-VTLN.

7.7 Zusammenfassung

Es wurden in diesem Kapitel Verfahren vorgestellt, die basierend auf den Daten einer einzigen Äußerung einen Parameter schätzen, der eine Sprechereigenschaft beschriebt, und mit dem die Muster durch Frequenzverzerrung bezüglich des Sprechers normiert werden können. Einige Nachteile der ML-Bestimmung des Verzerrungsfaktors im Training konnten durch das LD-Kriterium behoben werden und führten auch zu besseren Ergebnissen bei der Erkennung.

Im Vergleich zum Basiserkenner ohne VTLN fallen die Ergebnisse für leicht gestörte Kategorien besser aus, da die Sprechernormierung auch zu einer höheren Geräuschrobustheit beiträgt. Der Preis dafür ist ein zusätzlicher Erkennungslauf, mit dem über eine erste Hypothese eine Zuordnung der Muster zu den phonetischen Modellen des Erkenners gefunden wird.

Bei stärker gestörten Sprachdaten ist diese erste Hypothese sehr unzuverlässig. Die ML-Bewertung für die verschiedenen Frequenzverzerrungen der verrauschten Mustern ist ebenfalls nicht mehr so zuverlässig, denn sie beruht auf einem Modell für ungestörte Sprache. Daher verlieren die VTLN-Systeme hier gegenüber dem Basissystem bezüglich der Erkennungsleistung. Auf die Sprechernormierung sollte daher entweder in einigen Fällen (zum Beispiel unterhalb einer gewissen SNR-Schwelle) verzichtet werden oder sie müsste mit einem Geräuschreduktionsverfahren kombiniert werden. Eine weitere Möglichkeit wird in Kapitel 9 untersucht. Zur Bestimmung der Verzerrungsfaktoren wird dabei ein sprechernormiertes Hilfsmodell verwendet, das erstens keine phonetische Zuordnung benötigt und zweitens an das aktuelle Rauschen angepasst wird.

8 Kanal- und Rauschkompensation

„We call it noise
because we know nothing about it!"
– Bishnu Atal -

Kapitelübersicht: Wie im vorigen Kapitel werden hier Einflüsse betrachtet, die sich im Sprachsignal niederschlagen, aber nicht zur Erkennung beitragen und sich deshalb störend auswirken. Rauschen und wechselnde Kanäle werden zwar zunächst auch einzeln behandelt, aber aufgrund der starken Beeinflussung (besonders bei der Kanalschätzung) gemeinsam in diesem Kapitel betrachtet. Nach einer prinzipiellen Überlegung zur Kompensation von Kanal und Rauschen in 8.1 führt Abschnitt 8.2 ein Umgebungsmodell ein, das die beiden Störungen mathematisch beschreibt. Die Auswirkungen auf die in der Spracherkennung üblichen Merkmale analysiert Abschnitt 8.3. In Abschnitt 8.4 werden Geräuschreduktionsfilter untersucht, wie sie bereits erfolgreich für Kommandoworterkenner und zum Teil auch für kontinuierliche Erkenner eingesetzt wurden. Abschnitt 8.4.4 beschäftigt sich mit der Kanalkompensation und erweitert das Standardverfahren für den Einsatz in geräuschbehafteter Umgebung.

8.1 Prinzipielle Anwendung

Im Gegensatz zu robusten Merkmalen, bei denen die Störungen passiv unterdrückt werden, versuchen Kompensationsverfahren die Störungen aktiv zu reduzieren oder wechselnde Einflüsse zu normieren. Aus einer Modellvorstellung, die den Zusammenhang mit der Störung beschreibt, wird eine Kompensationsfunktion gefunden, deren Parameter bestimmt werden müssen. Die Anwendung besteht also meist aus zwei Schritten:

1. Schätzung der Störung (Gewinnung der Parameter).
2. Reduktion/Normierung durch geeignete Transformation des Signals.

Zu 1: In den seltensten Fällen kann die Störung direkt bestimmt werden, da sie sich mit anderen Einflüssen im Sprachsignal überlagert. Im Auto liegt solch ein Sonderfall vor, wenn man versucht die vom Lautsprecher (Radio, CD-Spieler usw.) erzeugten Überlagerungen in einer Sprachaufnahme zu reduzieren[16]. Das verursachende Störsignal ist dabei bekannt, nur nicht der Übertragungskanal über den es in die Aufnahme eingeht. Aber auch dieser kann

[16] Zur Vermeidung der Störursache (vergleiche Abschnitt 5.1) könnte man die Lautsprecherausgabe während einer Spracheingabe abschalten. Benutzerstudien im VODIS-Projekt zeigten jedoch, dass dies unangenehm auf den Benutzer wirkt. Um die Störreduktion dennoch etwas zu vereinfachen, kann man während einer Spracheingabe von Stereo- auf Monowiedergabe umschalten.

durch Ausgabe einer Zufallssequenz über den Lautsprecher zunächst näherungsweise bestimmt werden. Eine effektivere Reduktion wird über ein adaptives Echokompensationsverfahren erreicht. Diese Art der Störung durch bekannte Signale wird hier nicht weiter betrachtet. Das bedeutet, dass die Parameter der Störungen aus dem Sprachsignal geschätzt werden müssen.

Zu 2: Während es sich bei der Kompensation des Kanals praktisch immer um eine Normierung handelt, wird beim Rauschen meist eine Reduktion angestrebt. Statt zu versuchen den Einfluss des Rauschens komplett zu entfernen, kann man allerdings auch in diesem Fall durch eine „Rauschnormierung" dem Ziel der robusten Erkennung näher kommen. Indem man sowohl beim Training als auch bei der Erkennung die Daten durch Zumischen von Rauschen auf ein bestimmtes SNR-Niveau bringt, gleicht man die beiden Bedingungen an [Claes und Compernolle 1996, Gelin und Junqua 1999]. Dadurch sorgt man dafür, dass die Modelle auch bei leicht gestörten Situationen zu den Daten passen. Dies gilt aber nur solange bis das Rauschen während der Erkennung nicht den gewählten SNR-Wert unterschreitet. Außerdem verschlechtert man damit auch meist die Erkennung ungestörter Äußerungen. In [Song *et al.* 1998] wurden daher mehrere SNR-Niveaus verwendet, was aber auch bedeutet, dass man für jeden Fall ein eigenes akustisches Modell benötigt. Einen ähnlichen Weg beschreibt [Matrouf und Gauvain 1997]. Dort werden die Modelle mit dem Rauschen der zu erkennenden Äußerung kombiniert, während die Äußerung mit dem Rauschen der Trainingsdaten gemischt wird (gleichzeitige Modellanpassung und Normierung).

In der vorliegenden Arbeit wurde der Weg der Geräuschreduktion bevorzugt, da so die Varianz der Modelle klein gehalten wird, um auch für ungestörte Äußerungen gute Ergebnisse zu erzielen. Auf diese Weise ist es auch prinzipiell möglich ohne ein Neutraining auszukommen.

8.2 Modell des Übertragungskanals mit Rauschen

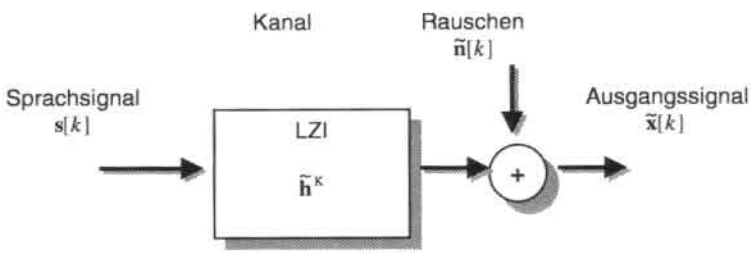


Abbildung 8.1: Kanalmodell für spektrale Merkmale.

Abbildung 8.1 zeigt ein gebräuchliches Kanalmodell, das als Grundlage der hier behandelten Verfahren verwendet wird. Die Übertragungseigenschaften $\tilde{\mathbf{h}}^{\kappa}$, bedingt durch

Raumakustik, Mikrofon und Aufnahmehardware, werden für die Dauer einer Aufnahme als linear und zeitinvariant betrachtet (LZI). Das bedeutet, dass ein Zeitsignal $s(t)$ mit der festen Impulsantwort $\tilde{h}^{K}(t)$ des Übertragungskanals gefaltet wird. Hintergrundgeräusche (wie auch zum Teil elektrische Störungen in der Aufnahmehardware) überlagern sich additiv. Nichtlineare Effekte, die zum Beispiel bei Übersteuerung entstehen können, werden hier vernachlässigt.

Betrachtet man die Fouriertransformierten der Signale (Spektren), so ergibt sich aus der Faltung eine Multiplikation, die Addition bleibt erhalten:

$$\text{Zeitbereich:} \quad x(t) = \tilde{h}^{K}(t) * s(t) + \tilde{n}(t) \tag{8.1}$$

$$\text{Spektralbereich:} \quad X(\omega) = \tilde{H}^{K}(\omega) \cdot S(\omega) + \tilde{N}(\omega) \tag{8.2}$$

Zur Verarbeitung im Spracherkenner werden aus dem aufgenommenen Signal Merkmalsvektoren für zeitlich aufeinanderfolgende Sprachrahmen k bestimmt, die aus einzelnen Koeffizienten bestehen:

$$\tilde{\mathbf{x}}[k] = \begin{pmatrix} \vdots \\ \tilde{x}_i[k] \\ \vdots \end{pmatrix}$$

Der Index i entspricht bei einem Kurzzeit-Spektrum (berechnet durch die diskrete Fouriertransformation, Gleichung (2.7)) bei N Abtastwerten und der Abtastfrequenz f_A, einer Analysefrequenz von

$$f = \frac{i}{N} f_A .$$

Für das Kurzzeit-Leistungsspektrum (Gleichung (2.9)) werden die Betragsquadrate der komplexwertigen Kurzzeit-Spektralkoeffizienten berechnet. Da in vielen Fällen das Sprachsignal nicht mit dem Rauschen korreliert ist und die Kanaleigenschaften sich nicht über die Zeit ändern, kommt man für dieses Merkmal zur folgenden Modellannahme:

$$\text{Leistungsspektrum:} \quad \tilde{x}_i[k] = \tilde{h}_i^{K} \cdot s_i[k] + \tilde{n}_i[k] . \tag{8.3}$$

Ist die Hintergrundstörung selbst sprachlicher Natur, trifft die Voraussetzung der unkorrelierten Störung nicht mehr zu, was aber hier nicht betrachtet wird.

Bei der Zusammenfassung der Leistungskoeffizienten zu *Mel*-Koeffizienten (mit einer *Mel*-Filterbank) bleibt der Zusammenhang aus Gleichung (8.3) auch für die resultierenden Merkmale bestehen. Auch für das Kurzzeit-Betragsspektrum (Gleichung (2.8)) wird manchmal dieser Zusammenhang verwendet, welcher sich für diesen Fall eigentlich nur dann ergibt, wenn die Phasen von komplexem Rausch- und Sprachspektrum übereinstimmen.

Vernachlässigt man den Einfluss des Rauschens und betrachtet logarithmierte Spektral- oder *Mel*-Koeffizienten, erhält man einen additiven Zusammenhang zwischen Kanal und Sprache:

$$\log(\tilde{x}_i[k]) = \log(\tilde{h}_i^{K}) + \log(s_i[k]) \tag{8.4}$$

Zur Vereinfachung der Darstellung benutzen wir nun wieder komplette Merkmalsvektoren und schreiben logarithmierte, spektrale Merkmale („*Log*-Spektrum") mit einem hochgestellten „l" und cepstrale Merkmale mit hochgestelltem „c". Da cepstrale Merkmale aus einer linearen Abbildung von *Log*-Spektralmerkmalen hervorgehen, gilt der additive Zusammenhang mit dem Kanal auch für diese:

Log-Spektrum $$\tilde{\mathbf{x}}^{l}[k] = \tilde{\mathbf{h}}^{\kappa l} + \mathbf{s}^{l}[k] \tag{8.5}$$

Cepstrum $$\tilde{\mathbf{x}}^{c}[k] = \tilde{\mathbf{h}}^{\kappa c} + \mathbf{s}^{c}[k] \tag{8.6}$$

Es soll nun gezeigt werden, wie sich prinzipiell der Kanal aus dem Signal extrahieren lässt. Dies wird exemplarisch für das Cepstrum gezeigt, kann aber auch analog dazu auf dem *Log*-Spektrum basieren.

Da der Kanal als zeitinvariant angenommen wurde, bleibt der Beitrag auch nach einer Mittelwertbildung über die Zeit erhalten:

$$\frac{1}{K}\sum_{k=1}^{K}\tilde{\mathbf{x}}^{c}[k] = \tilde{\mathbf{h}}^{\kappa c} + \frac{1}{K}\sum_{k=1}^{K}\mathbf{s}^{c}[k] \tag{8.7}$$

In manchen Veröffentlichungen wird angenommen, dass die rechte Summe in Gleichung (8.7) verschwindet, was so nicht stimmt. Für die Spracherkennung ist sie aber meist nicht interessant, da sie hauptsächlich vom Sprecher und der Verteilung der geäußerten Laute abhängt. Sind Sprecher und Verteilung der Laute über die jeweils betrachteten Stichproben konstant, können wir die Summe als **Sprecherkanal** $\tilde{\mathbf{h}}^{sc}$ bezeichnen[17] und mit $\tilde{\mathbf{h}}^{\kappa c}$ zusammenziehen:

$$\frac{1}{K}\sum_{k=1}^{K}\tilde{\mathbf{x}}^{c}[k] = \tilde{\mathbf{h}}^{\kappa c} + \tilde{\mathbf{h}}^{sc} = \tilde{\mathbf{h}}^{c} \tag{8.8}$$

Subtrahiert man den resultierenden Kanal $\tilde{\mathbf{h}}^{c}$ von $\tilde{\mathbf{x}}^{c}[k]$, bleibt nur der mittelwertfreie, lautinformationstragende Sprachanteil übrig. Dieser ist nun unabhängig vom Übertragungskanal. Dieses sehr einfache und effektive Kanalkompensationsverfahren nennt sich

Cepstrale Mittelwertsubtraktion (CMS): $$\hat{\mathbf{s}}^{c}[k] = \tilde{\mathbf{x}}^{c}[k] - \frac{1}{K}\sum_{k=1}^{K}\tilde{\mathbf{x}}^{c}[k] \tag{8.9}$$

und wird in Abschnitt 8.5 noch ausführlich behandelt.

[17] Was hier als Sprecherkanal bezeichnet wird, entspricht dem Langzeitspektrum (S. 108) im Spektralbereich. Bedingt durch die Mittelung über unterschiedliche Arten von Merkmalen kann man sie jedoch nicht direkt in Beziehung miteinander setzen.

8.3 Auswirkung auf die Merkmale

8.3.1 Analyse der gemittelten Spektren

Abbildung 8.2 zeigt die gemittelten Spektren (siehe Seite 42) dreier vom Wortlaut gleicher Äußerungen eines Sprechers, aufgenommen bei unterschiedlich starkem Hintergrundgeräusch. Der dargestellte Frequenzbereich entspricht 0 bis 8000 Hz. Deutlich ist erkennbar, wie die Intensität des Rauschens steigt und sich die Bereiche der Standardabweichungen (dünne Linien) von Rauschen und Sprache immer weiter überlagern. In der mittleren und der rechten Abbildung sind auch besonders ausgeprägte Geräuschspitzen bei bestimmten Frequenzen zu erkennen (markiert durch Pfeile).

Bei stärkerem Geräusch fällt auf, dass auch das Sprachspektrum größere Werte annimmt. Der Sprecher lässt sich durch die Umgebung beeinflussen (Lombard-Effekt). Die Änderungen bewirken vor allem eine Verschiebung des logarithmierten Sprachspektrums, die hier relativ gleichmäßig über alle Frequenzen ausfällt. Das entspricht einer unterschiedlichen Lautstärke des Sprachsignals, was aber beispielsweise durch die Cepstrale Mittelwertsubtraktion ausgeglichen werden kann.

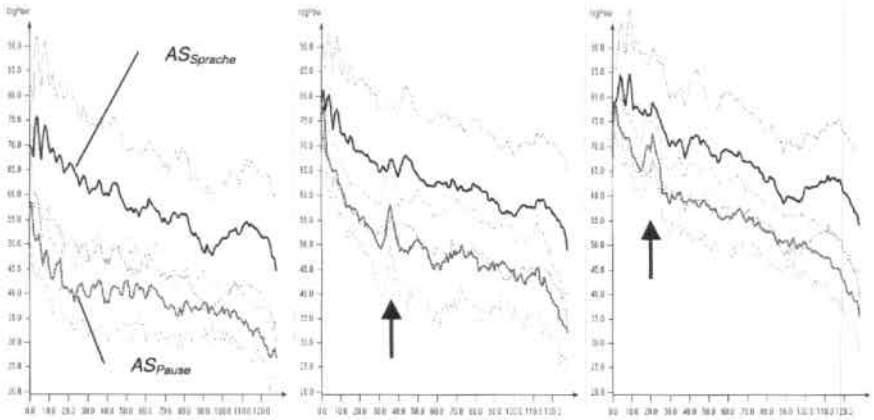


Abbildung 8.2: Gemittelte Spektren dreier Aufnahmen:
Sprachspektrum ($AS_{Sprache}$, oben), Pausenspektrum (AS_{Pause}, unten).
Die dünnen Linien deuten die jeweiligen Standardabweichungen nach oben und unten an.

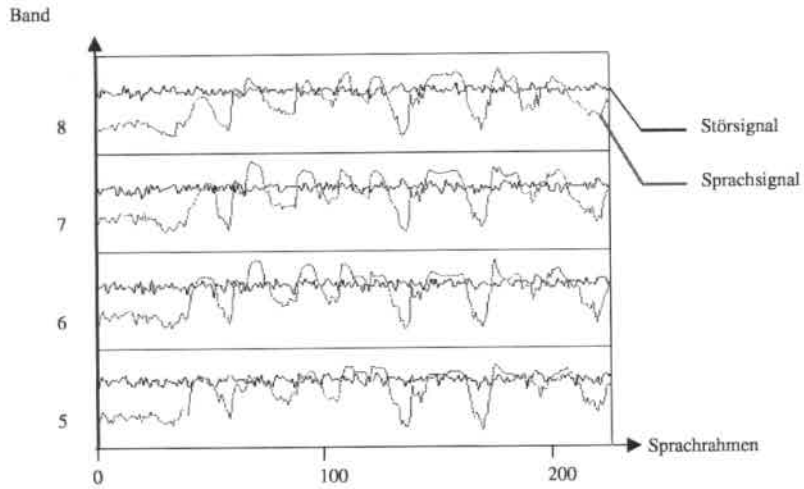


Abbildung 8.3: Der zeitliche Verlauf von vier logarithmierten *Mel*-Bändern eines ungestörten Sprachsignals und eines Störsignal (Fahrgeräusch 100km/h).

8.3.2 Zeitlicher Verlauf, „Missing Features"

Die **Abbildung 8.3** stellt den zeitlichen Verlauf von *Log*-Spektralmerkmalen für Sprache und Fahrgeräusch dar. Im Vergleich zum Fahrgeräusch weist die Sprache für diese Merkmale große Schwankungen auf. Für das Beispiel haben die relativ flach verlaufenden Kurven des Rauschens die mittlere Intensität der Sprache überschritten. Werden die Signale nun im linearen Bereich addiert und anschließend logarithmiert, ergibt sich ein Verlauf, der in etwa dem jeweiligen Maximum von Sprache und Rauschen entspricht.

Dieses Verhalten wird auch durch Betrachtung der Histogramme eines solchen Merkmals deutlich (**Abbildung 8.4** bis **Abbildung 8.7**). Die ersten beiden Abbildungen zeigen die relativen Häufigkeiten bestimmter Wertebereiche eines logarithmierten *Mel*-Spektralkoeffizienten von Sprache und Fahrgeräusch über die Dauer einer Äußerung. Die Werte des ungestörten Sprachsignals verteilen sich fast gleichmäßig über einen großen Bereich, während die Werte für die Geräuschaufnahme sich relativ dicht um einen mittleren Wert gruppieren. Die Form dieser Verteilung hat starke Ähnlichkeit mit einer Gaußdichte, weshalb diese Näherung auch im folgenden Kapitel für Rauschsignale verwendet wird. Die beiden restlichen Abbildungen zeigen die resultierende Verteilung nachdem das Rauschen (im Zeitbereich) gewichtet zum Sprachsignal addiert wurde. Die Verteilung des gestörten Sprachsignals wird für große Bereiche vom Rauschen dominiert.

Die kleinen Werte des Sprachsignals werden durch starkes Rauschen überdeckt und ihr Informationsgehalt geht damit verloren. An gewissen Stellen fehlen also die Merkmale des Sprachsignals (engl.: *missing features*). Ein neuerer Trend in der Spracherkennung versucht diesen Umstand in der *Missing Feature Theory* zu berücksichtigen (siehe Abschnitt 5.2).

Abbildung 8.4 bis **Abbildung 8.7:** Histogramme (relative Häufigkeiten h quantisierter Werte x) der logarithmierten *Mel*-Filterbankwerte (Band 5) für eine Äußerung.

Abbildung 8.4: ungestörtes Sprachsignal.

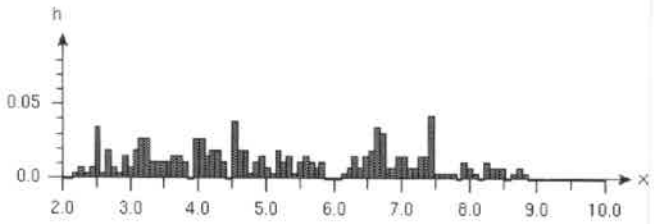


Abbildung 8.5: Fahrgeräusch 100km/h.

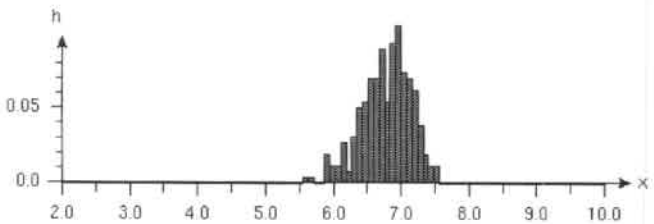


Abbildung 8.6: Sprachsignal mit Fahrgeräusch .

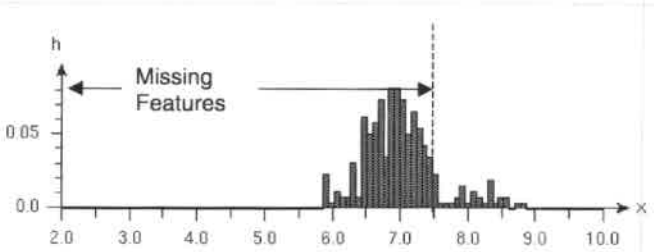
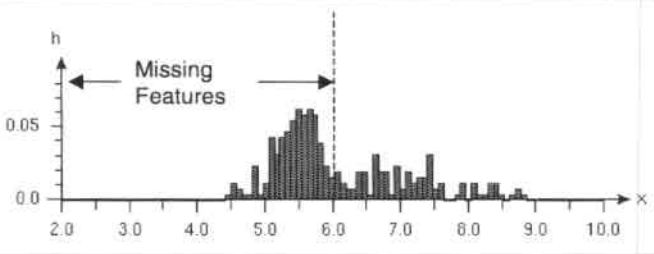


Abbildung 8.7: Sprachsignal mit Fahrgeräusch (Gewichtung 0,2).
Die Gewichtung entspricht einer Verschiebung des Fahrgeräusches um $\log(0{,}2^2) \approx -1{,}4$ entlang der x-Achse.



8.3.3 Veränderung der Merkmale durch additives Rauschen

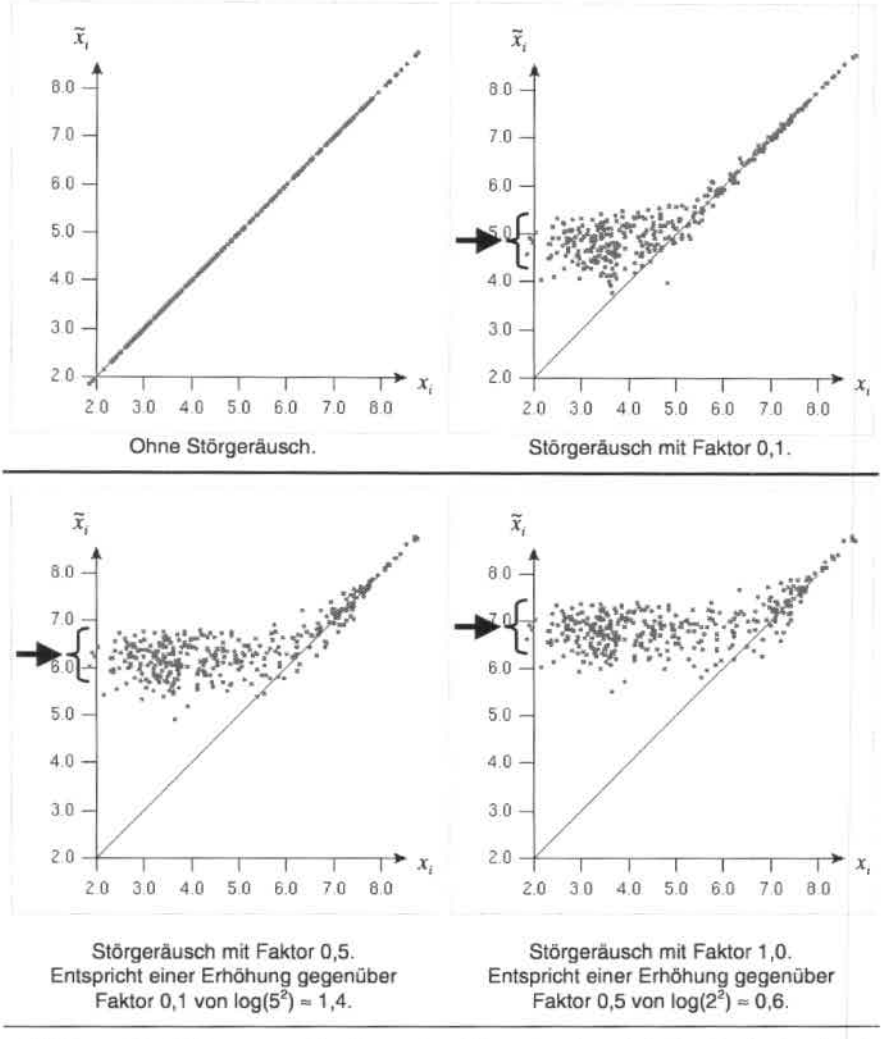
Die vorigen Abschnitte haben sich eher mit den generellen Eigenschaften und Auswirkungen wechselnder Kanäle und Geräusche beschäftigt. Für das in der Spracherkennung wichtige *Log*-Spetrum und das Cepstrum gilt ein einfacher Zusammenhang zwischen Sprachanteil und Kanal, der sich in einer konstanten Verschiebung aller Werte ausdrückt. Für additives Rauschen, das in der Fahrzeugumgebung eine besonders wichtige Rolle spielt, sind die Zusammenhänge nicht so einfach. Daher soll nun betrachtet werden, wie sich diese Art der Störung auf einzelne Merkmalswerte auswirkt.

In **Abbildung 8.8** sind alle Werte eines *Log*-Spektralmerkmals einer Äußerung als Punkte eingetragen. Die Abszisse (x_i) gibt den Wert des ungestörten Merkmals an, während die Ordinate ($\tilde{x}_i$) das Ergebnis nach Zumischung eines Rauschsignals darstellt. Sind die Merkmale ungestört, so befinden sich alle Punkte auf der Winkelhalbierenden. Wird Rauschen zugemischt, weichen bestimmte Punkte immer mehr von ihrem ungestörten Pendant ab und entfernen sich von der Winkelhalbierenden. Dies betrifft, wie oben schon festgestellt, vor allem die kleinen Werte, die sich nun nicht mehr unterscheiden lassen. Die mittlere Stärke des Rauschens ist jeweils mit einem Pfeil und seine Schwankungen mit einer Klammer angedeutet.

Was bedeutet dies für die Spracherkennung? Wurde ein Erkenner mit ungestörten Merkmalen trainiert, so verursachen diese Abweichungen Fehleinschätzungen der Klassenzugehörigkeit, weshalb man versucht, den Unterschied zum ungestörten Merkmal zu minimieren. Ein Kriterium, das dazu verwendet werden kann, ist der mittlere quadratische Fehler.

Mithilfe des Kanalmodells und einer Schätzung des Rauschmittelwertes ist es möglich diesen Fehler zu reduzieren. Hier setzen eine Reihe von Verfahren an, deren bekanntester Vertreter die **Spektrale Subtraktion** ist ([Boll 1979], siehe nächster Abschnitt). Bei dieser wird angenommen das Rauschen habe eine verschwindend kleine Varianz. Da nach dem Kanalmodell Störung und Sprache sich im linearen Spektralbereich additiv verhalten, wird bei diesem Verfahren einfach das geschätzte Rauschen vom gestörten Signal subtrahiert (siehe **Abbildung 8.9**).

Liegt für eine bestimmte Art von Störung eine Trainingsmenge mit jeweils gestörten und ungestörten Merkmalen vor (oder erzeugt man sich diese wie hier durch Zumischen), kann man eine Abbildungsvorschrift auch ohne ein Modell der Störung finden. Unter Umständen können mit diesen „Stereodaten" auch nichtlineare Zusammenhänge berücksichtigt werden. In [Tamura und Waibel 1988] und [Trompf 1996] wird diese Abbildung für verrauschte Merkmale mithilfe von Neuronalen Netzen gelernt. Voraussetzung ist, wie bereits erwähnt, eine Trainingsmenge mit zwei Versionen der Merkmale, was im vorliegenden Fall nicht für alle Störfaktoren gegeben ist. Außerdem wäre eine so gelernte Abbildung nur für ganz bestimmte Störbedingungen verwendbar, nämlich solche, die auch in der Trainingsmenge präsent waren.



Ohne Störgeräusch.

Störgeräusch mit Faktor 0,1.

Störgeräusch mit Faktor 0,5.
Entspricht einer Erhöhung gegenüber
Faktor 0,1 von $\log(5^2) \approx 1{,}4$.

Störgeräusch mit Faktor 1,0.
Entspricht einer Erhöhung gegenüber
Faktor 0,5 von $\log(2^2) \approx 0{,}6$.

Abbildung 8.8: Einfluss unterschiedlich starker Störgeräusche auf ein *Log*-Spektralmerkmal (Logarithmus des fünften *Mel*-Bandes).

Die Punkte entsprechen den Mustern dieses Merkmals für eine Äußerung:

- x_i ungestört und
- $\tilde{x}_i$ mit zugemischtem Rauschen (Fahrgeräusch bei 100 km/h mit verschiedenen Mischungsfaktoren).

Die Pfeile geben die ungefähre Intensität des Rauschens an,
die Klammern dessen ungefähre Schwankungen.

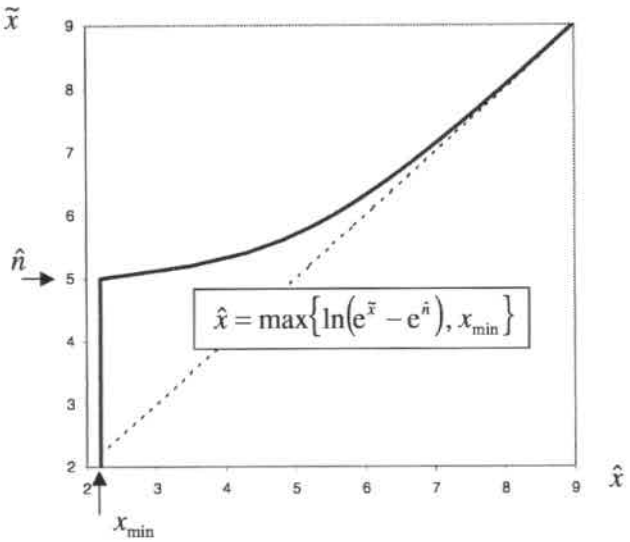


Abbildung 8.9: Durch Spektrale Subtraktion mit einem Rauschmittelwert wird der gestörte Merkmalswert $\tilde{x}_i$ auf einen Schätzwert $\hat{x}_i$ (für den ungestörten Wert x_i) abgebildet. Die Funktion ist hier etwas unkonventionell über die Ordinate aufgetragen, damit der Zusammenhang mit den vorigen Abbildungen deutlicher wird.

8.4 Geräuschreduktionsfilter

Dieser Abschnitt beschreibt die Geräuschreduktion nach dem Verfahren der Spektralen Subtraktion und ein weiteres Geräuschreduktionsfilter nach Ephraim und Malah. Besonders die Spektrale Subtraktion wird häufig bei Einzelworterkennern, seltener auch bei kontinuierlichen Erkennern eingesetzt. Beide Verfahren wurden in dieser Arbeit auch experimentell untersucht.

8.4.1 Spektrale Subtraktion

Die Spektrale Subtraktion ist eine Technik zur Geräuschreduktion, die den Einfluss additiver Störungen vermindern soll. Obwohl [Boll 1979] sie bereits zum Einsatz für die Spracherkennung vorgeschlagen hat, wurde sie ursprünglich zur Verbesserung der Sprachqualität in der Übertragungstechnik entwickelt. Als Voraussetzung wird angenommen, dass die Störung, die das Signal rein additiv überlagert, auch unkorreliert mit dem Sprachsignal ist. Es handelt sich also um einen hier schon behandelten Spezialfall des oben eingeführten Kanalmodells.

Ein Schätzwert für das Rauschen $\hat{\mathbf{n}}[k]$ wird meist während der Sprachpausen durch Mittelung der gestörten Merkmalsvektoren $\tilde{\mathbf{x}}[k]$ berechnet und hängt dann nicht mehr vom Zeitindex k ab. Dafür ist ein Sprach-Pause-Detektor nötig. Außerdem wird damit zusätzlich angenommen, dass die Störung quasi stationär ist und somit die Schwankungen der berechneten Geräuschstatistik über die Zeit klein bleiben.

Die ursprüngliche Definition der Spektralen Subtraktion für Kurzzeit-Betragsspektren (angedeutet durch ein hochgestelltes m) zur Berechnung eines Schätzwertes für ungestörte Sprache $\hat{\mathbf{x}}[k]$ lautet:

$$\hat{\mathbf{x}}^{\mathrm{m}}[k] = \tilde{\mathbf{x}}^{\mathrm{m}}[k] - \hat{\mathbf{n}}^{\mathrm{m}}[k] \tag{8.10}$$

Zur Rekonstruktion des Sprachsignals in der Übertragungstechnik wird die Phase vom gestörten Kurzzeitspektrum übernommen und mit der sogenannten *„overlap-add"*-Methode (siehe z.B. [Kammeyer und Kroschel 1989, S.182]) ein Zeitsignal rekonstruiert. Für die Verwendung in einem Spracherkenner ist dies nicht erforderlich, da die gängigen Merkmalsextraktionen direkt auf dem Betrags- bzw. Leistungsspektrum aufbauen. Wird dennoch die Spektrale Subtraktion als Filter mit den Verstärkungsfaktoren $g_i[k]$ für das komplexwertige Spektrum $\tilde{\mathbf{x}}[k]$ betrachtet, ergibt sich:

$$\hat{x}_i[k] = g_i[k] \cdot \tilde{x}_i[k] \quad \text{mit} \quad g_i[k] = \frac{\tilde{x}_i^{\mathrm{m}}[k] - \hat{n}_i^{\mathrm{m}}[k]}{y_i^{\mathrm{m}}[k]} \tag{8.11}$$

Wird hier statt der Beträge das Leistungsspektrum verwendet, so entspricht die Spektrale Subtraktion einem adaptiven Wiener-Filter (siehe z.B. [Kroschel 1988, Carstens-Behrens 1996]). Verallgemeinert man diese Filterdefinition, können damit verschiedene Realisierungsformen der Spektralen Subtraktion beschrieben werden:

$$g_i[k] = \left(\frac{\left(\tilde{x}_i^{\mathrm{m}}[k]\right)^{\gamma} - \left(\hat{n}_i^{\mathrm{m}}[k]\right)^{\gamma}}{\left(\tilde{x}_i^{\mathrm{m}}[k]\right)^{\gamma}} \right)^{\delta} \tag{8.12}$$

Für $\gamma = \delta = 1$ erhält man die Subtraktion der Beträge, mit $\gamma = 2, \delta = 1/2$ die Subtraktion der Leistungsspektren und mit $\gamma = 2, \delta = 1$ das Wiener-Filter.

Die effiziente Umsetzbarkeit (geringer Rechenzeitbedarf) erklärt die weite Verbreitung, wobei die Effektivität (Auswirkung auf die Erkennungsleistung) unterschiedlich beurteilt wird. Letztere hängt stark von der Art des Rauschens und der Rauschschätzung ab (siehe dazu [Hirsch und Ehrlicher 1995]). Offensichtliche Nachteile werden in erweiterten Formen mehr oder minder gut kompensiert:

- Durch die Mittelung bei der Rauschschätzung bleiben nach der Subtraktion Reste in einigen Spektralbändern übrig. Diese werden bei der Rekonstruktion als *„musical tones"* wahrgenommen und stören nicht nur den Höreindruck sondern auch die Erkennung.
- Ist das geschätzte Rauschen in einem Band größer als der aktuelle Signalwert würden negative Werte für das Betragsspektrum entstehen und müssen zumindest auf Null gesetzt werden.

In [Berouti *et al.* 1979] wird daher ein Überschätzfaktor $\alpha \geq 1$ (engl.: *over estimation factor*) und ein Restrauschfaktor $0 \leq \beta \leq 1$ (engl.: *spectral floor*) eingeführt. Dort wird mit der Subtraktion des Leistungsspektrums gearbeitet, das folgend vereinfacht als Merkmalsvektor mit hochgestelltem „s" geschrieben wird:

$$\hat{x}_i^s[k] = \max\left\{\tilde{x}_i^s[k] - \alpha \cdot \hat{n}_i^s[k],\ \beta \cdot \hat{n}_i^s[k]\right\} \tag{8.13}$$

Die Leistungsspitzen des Restrauschens werden durch die Überschätzung gedämpft und die Täler durch das Restrauschen überdeckt. Somit werden die *„musical tones"* reduziert, was je nach Rauschen durch geeignete Wahl der Faktoren optimiert werden muss. Dies kann auch abhängig vom augenblicklichen SNR geschehen, wofür nochmals zusätzliche Parameter nötig sind. Eine Erweiterung dieser Art ist die nichtlineare Spektrale Subtraktion von [Lockwood und Boudy 1992].

Die Frage, auf welcher Art von Merkmalen die Spektrale Subtraktion anzuwenden ist, kann nicht eindeutig beantwortet werden, wird meist jedoch zugunsten des Leistungsspektrums entschieden. Vorteile ergeben sich auch bei der Anwendung auf zusammengefasste Spektralbänder, wie dies beispielsweise durch die *Mel*-Filterbank geschieht. Für diesen Fall kann die Spektrale Subtraktion auch auf die logarithmierten Merkmalskoeffizienten umgerechnet werden, wie dies bereits in **Abbildung 8.9** angedeutet wurde. In [Acero 1990] und [Compernolle 1989] wird diese Art der Umsetzung beschrieben. Als Vorteil ergibt sich, dass man in einem Merkmalsraum operiert, in dem die Distanz zwischen Sprachmustern wesentlich aussagekräftiger ist als im linearen Spektralbereich [Gray *et al.* 1976]. Als Optimierungskriterium zum Einstellen der freien Parameter kann man dann zum Beispiel den mittleren quadratischen Fehler (MSE) benutzen, sofern die ungestörten Merkmale vorliegen (s.u., S. 145).

Neben den Parametern für die Geräuschschätzung müssen vor allem Werte für den Überschätzfaktor α und den Restrauschfaktor β gefunden werden. **Abbildung 8.10** und **Abbildung 8.11** zeigen die Auswirkungen des Rauschens und der Geräuschreduktion auf einen logarithmierten *Mel*-Koeffizienten. In der hier verwendeten Implementierung wurde jedoch abweichend von der Gleichung (8.13) für das Restrauschen $\beta \cdot \tilde{x}_i^s[k]$ verwendet, wodurch sich keine allzu abrupte Begrenzung ergibt. Durch die additive Störung weichen besonders die kleinen Werte vom ungestörten Merkmal ab, die Entfernung zur Winkelhalbierenden ist entsprechend groß. Bei dieser Art der Darstellung würde die ideale Geräuschreduktion alle abweichenden Punkte (gestörte Muster) wieder auf die Winkelhalbierende abbilden. Damit entspricht der Schätzwert der Geräuschreduktion dem ungestörten Wert und der Fehler verschwindet.

Mit dem Überschätzfaktor α wird die Stärke der Reduktion bestimmt (**Abbildung 8.10**). Durch ihn werden die Schätzwerte im Bereich zwischen dem gestörten Wert und dem Restrauschen verteilt. Ist er zu klein wird kaum kompensiert, bei zu großen Werten oft durch das Restrauschen begrenzt. Im letzteren Fall werden auch häufiger wenig gestörte Merkmalswerte verändert.

Abbildung 8.11 zeigt den Einfluss des Restrauschfaktors β auf den geschätzten *Mel*-Koeffizienten. Ist β zu groß gewählt (links) verbleibt ein starkes Restrauschen. Bei zu kleinem β (rechts) werden viele Werte zu klein und vergrößern so ebenfalls den Fehler.

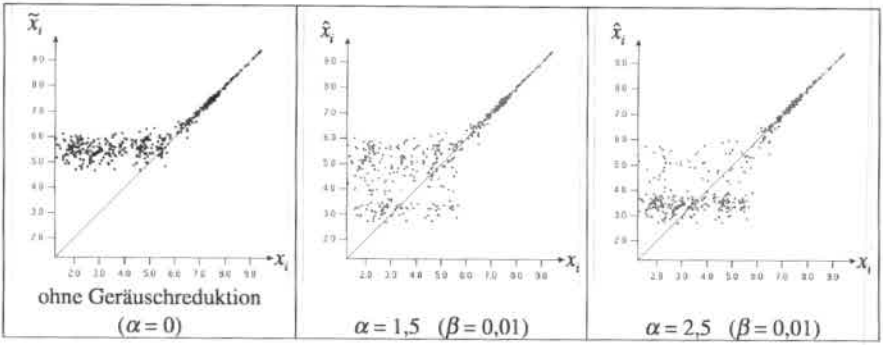


Abbildung 8.10: Muster eines *Log*-Spektralmerkmales für alle Sprachrahmen einer Äußerung als Punkte dargestellt. x_i ungestört und $\hat{x}_i$ mit zugemischtem Rauschen und Geräuschreduktion durch Spektrale Subtraktion (verschiedene Überschätzfaktoren α).

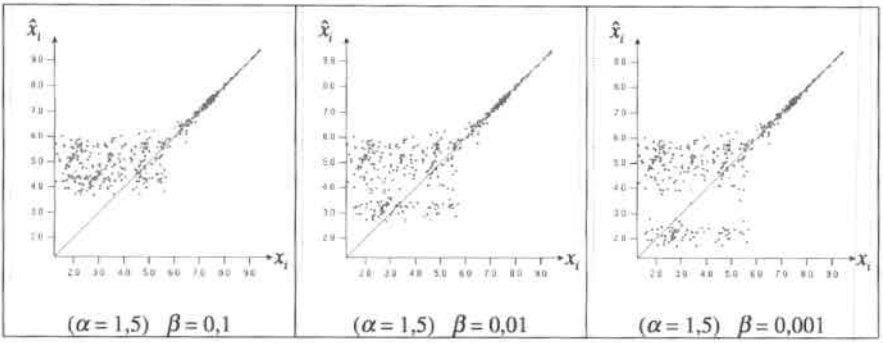


Abbildung 8.11: Muster eines *Log*-Spektralmerkmales für alle Sprachrahmen einer Äußerung als Punkte dargestellt. x_i ungestört und $\hat{x}_i$ mit zugemischtem Rauschen und Geräuschreduktion durch Spektrale Subtraktion (verschiedene Restrauschfaktoren β).

8.4.2 Geräuschreduktionsfilter nach Ephraim und Malah

Neben den oben beschriebenen und häufig benutzten Verfahren nach der Spektralen Subtraktion existieren noch einige weitere einkanalige Verfahren zur Geräuschreduktion. Einer der effektivsten Ansätze stammt von Ephraim und Malah und basiert direkt auf dem MSE-Kriterium. Wie in Abschnitt 3.3.2 erörtert, ergibt sich dafür der Schätzwert als Erwartungswert der bedingten Dichtefunktion $p(x_i \mid \tilde{x}_i)$.

Dies führt in diesem Fall zu recht komplexen mathematischen Ausdrücken für die es keine geschlossene Lösung gibt. Eine datengetriebene Lösung wurde daher in [Porter und Boll 1984] entwickelt. [Ephraim und Malah 1983, 1984] stellen eine geschlossene Lösung vor, die voraussetzt, dass die spektralen Merkmale normalverteilt und statistisch unabhängig sind. Der

Verstärkungsfaktor $g_i[k]$, mit dem sich der Schätzwert $\hat{x}_i[k]$ durch Multiplikation mit dem gestörten Merkmal $\tilde{x}_i[k]$ ergibt, lautet

$$g[k]=\frac{\sqrt{\pi}}{2}\sqrt{\left(\frac{1}{1+SNR_{post}[k]}\right)\cdot\left(\frac{SNR_{prio}[k]}{1+SNR_{prio}[k]}\right)}\cdot M\left[\left(1+SNR_{post}[k]\right)\cdot\left(\frac{SNR_{prio}[k]}{1+SNR_{prio}[k]}\right)\right] \quad (8.14)$$

mit $M[\theta]=e^{\frac{\theta}{2}}\left[(1+\theta)I_0\left(\frac{\theta}{2}\right)+\theta\, I_1\left(\frac{\theta}{2}\right)\right]$ und den Besselfunktionen I_0 und I_1.

Die Indizes i für die einzelnen Koeffizienten wurden zur Vereinfachung bei dem Verstärkungsfaktor und den beiden SNR-Werten weggelassen. Letztere sind die bestimmenden Einflussfaktoren und werden nun erläutert:

- Der A-priori-SNR, hier mit SNR_{prio} bezeichnet, ist der dominierende Faktor. Nur wenn er große Werte annimmt, ergibt sich ein Verstärkungsfaktor $g[k]$ nahe 1. Das heißt, für diesen Fall wird das (wenig gestörte) Signal unverändert gelassen.

$$SNR_{prio}[k]=\max\left\{(1-\eta)\cdot SNR_{post}[k]+\eta\cdot\frac{\hat{x}^2[k-1]}{\hat{n}^2[k-1]},\, SNR_{prio\,min}\right\} \quad (8.15)$$

- Der A-posteriori-SNR, hier als SNR_{post} bezeichnet, ist der augenblickliche Signal-zu-Rausch-Wert und ergibt sich direkt aus dem Verhältnis von $\tilde{x}_i^2[k]-\hat{n}_i^2[k]$ und dem geschätzten Rauschen $\hat{n}_i^2[k]$. Er wirkt sich erst aus, wenn SNR_{prio} kleinere Werte annimmt, d.h. die Signalstärke über mehrere Analyserahmen im Verhältnis zur Störung abnimmt. Für einen großen lokalen Wert SNR_{post} ist dann der Verstärkungsfaktor $g[k]$ besonders klein, was zunächst nicht intuitiv erscheint.

$$SNR_{post}[k]=\max\left\{\frac{\tilde{x}^2[k]-\hat{n}^2[k]}{\hat{n}^2[k]},\, SNR_{post\,min}\right\} \quad (8.16)$$

Das Verhalten von SNR_{post} lässt sich folgendermaßen erklären: SNR_{prio} ist durch eine Glättung (mit beispielsweise $\eta=0{,}98$) aus den zeitlich vorangegangenen Werten von SNR_{post} entstanden und stellt insofern eine Unterscheidung in Sprache und Pause dar. In Sprachpausen (kleines SNR_{prio}) werden dann Abweichungen vom geschätzten Rauschen nach oben (großes SNR_{post}) stärker gedämpft. Dieses Verfahren nach Ephraim und Malah unterdrückt weitestgehend die „*musical tones*" und entspricht, gesteuert durch SNR_{post}, verschiedenen Ausprägungen der klassischen Spektralen Subtraktion. Eine genauere Analyse der Wirkungsweise und Beschreibung der Einzelheiten findet man in [Cappé 1994].

Abbildung 8.12 zeigt einen Vergleich mit der Spektralen Subtraktion wie sie oben erörtert wurde. Für beide Verfahren erfolgte die Geräuschschätzung über Minimum-Statistik [Martin 1993]. Dabei werden die Koeffizienten des Leistungsspektrums geglättet und über ein Zeitfenster das jeweilige Minimum bestimmt. Durch einen Überschätzfaktor ergibt sich ein Schätzwert $\hat{n}_i^2[k]$ für das Rauschen.

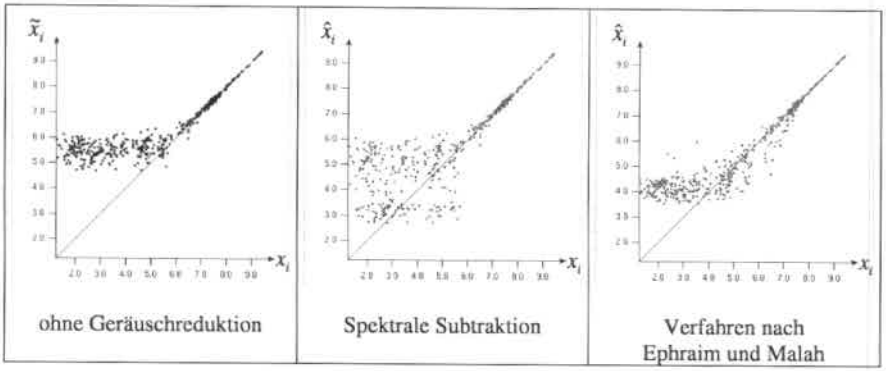


Abbildung 8.12: Ein Vergleich der beiden Geräuschreduktionsfilter.

8.4.3 Versuchsergebnisse

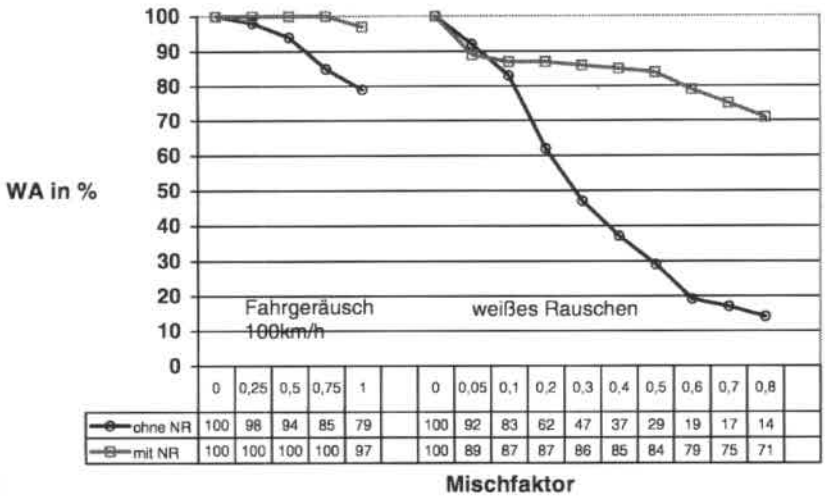
Durch den Einsatz der oben beschriebenen, einkanaligen Verfahren sollten vor allem die Störungen, die durch das starke Rauschen im Fahrzeug verursacht wurden, kompensiert werden. Aus der Literatur ist bekannt, dass die Spektrale Subtraktion für ungestörte Daten oft eine Verschlechterung der Erkennungsleistung ergibt, zumindest wenn nicht schon mit ihr trainiert wurde.

Beide vorgestellten Verfahren haben eine Reihe offener Parameter, zumal sie auch von der benötigten Rauschschätzung abhängen. Eine Einstellung der Parameter mit dem großen Optimierungskreislauf (**Abbildung 3.1**, S. 38) wie sie zum Beispiel in [Schless und Class 1998] für einen Ziffernerkenner durchgeführt wurde, kommt im vorliegenden Fall nicht in Frage. Der Aufwand für Training und Test des hier verwendeten kontinuierlichen Erkenners für mehrere Parameterkombinationen wäre zu groß.

Einzelworterkenner

In Vorversuchen mit kleineren und schnelleren Einzelworterkennern wurden daher die Parameter der beiden Verfahren eingestellt und die Ergebnisse verglichen. Dort zeigte sich das etwas aufwändigere Verfahren von Ephraim und Malah in Kombination mit einer Rauschschätzung mit Minimum-Statistik [Martin 1993] als überlegen.

Abbildung 8.13 zeigt die Resultate eines Erkenners für 13 Komandowörter auf Labordaten mit zugemischtem Rauschen (verschiedene Mischungsfaktoren). Durch Addieren des Fahrgeräusches sinkt die Erkennungsleistung ohne Geräuschreduktion von 100 % auf knapp 80 % ab. Mit der Geräuschreduktion bleibt die Erkennungsrate deutlich über 95 % und das, obwohl sich der Sprecher unter realistischen Verhältnissen in der Lautstärke anpassen würde, um damit das Rauschen etwas auszugleichen. Erst durch Zumischen von sehr starkem weißen Rauschen werden die Ergebnisse deutlich schlechter (mit SNR teilweise unter 0 dB). Auch für diesen Fall zeigt die Geräuschreduktion, wenn auch auf einem schlechteren Niveau, weit bessere Ergebnisse.



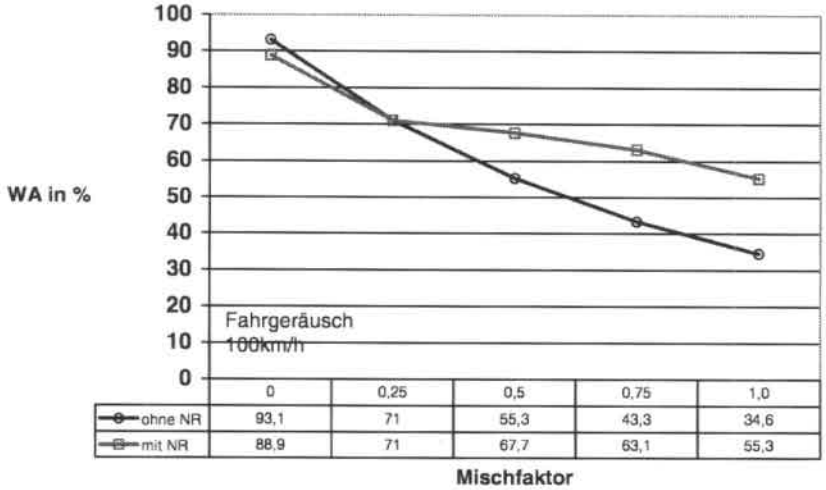
	0	0,25	0,5	0,75	1		0	0,05	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8
ohne NR	100	98	94	85	79		100	92	83	62	47	37	29	19	17	14
mit NR	100	100	100	100	97		100	89	87	87	86	85	84	79	75	71

Abbildung 8.13: Ergebnisse des Kommandoworterkenners mit und ohne Geräuschreduktion nach Ephraim und Malah.

Für diesen Anwendungsfall mit sehr kleinem Vokabular erhöht also die Geräuschreduktion durch das Filter die Zuverlässigkeit der Klassifikation. Wie sieht es aber bei größerem Vokabular aus, bei dem zum Teil sehr ähnliche Wörter vorkommen, die sich nur durch wenige Laute unterscheiden? Dazu wurde ein Testset mit potentiell allen 1.712 Straßennamen aus Karlsruhe verwendet. Die Ergebnisse in **Abbildung 8.14** sind erwartungsgemäß um einiges schlechter und die Erkennungsrate ohne Geräuschreduktion fällt bis unter 40 % ab. Mit dem Filter können zwar wiederum starke Verbesserungen erzielt werden, diese fallen aber nicht mehr in der Deutlichkeit aus wie bei den Kommandowörtern.

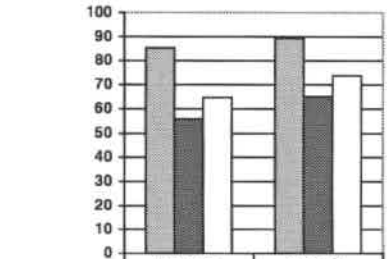
Kontinuierliche Erkennung mit künstlich verrauschten Daten

Bei der Erkennung kontinuierlicher Sprache mit einem noch größeren Vokabular verlor das Geräuschreduktionsfilter noch mehr an Wirksamkeit. Einen großen Einfluss hat dabei die Verwendung der Linearen Diskriminanzanalyse (LDA, Abschnitt 5.4.2, S. 81). Für einen Test auf den Navigationsanfragen mit zugemischtem Rauschen zeigt **Abbildung 8.15** einen Vergleich zweier Systeme, die ähnlich trainiert wurden wie das Basissystem. Das System ohne LDA (links) profitiert von der Geräuschreduktion, während bei dem System mit LDA (rechts) nur die Wortkorrektrate WC besser ausfällt. Die LDA trägt allerdings selbst stark zur Robustheit bei, wie man am Vergleich der Ergebnisse auf den gestörten Daten ohne Geräuschreduktion sieht: Das Ergebnis von 64,5 % ist fast 9 % absolut besser als bei dem System ohne LDA.



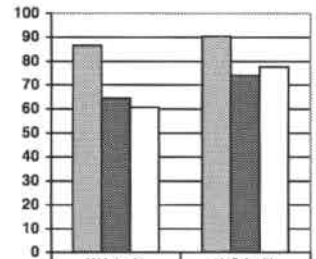
	0	0,25	0,5	0,75	1,0
ohne NR	93,1	71	55,3	43,3	34,6
mit NR	88,9	71	67,7	63,1	55,3

Abbildung 8.14: Ergebnisse des Straßennamen-Erkenners mit und ohne Geräuschreduktion nach Ephraim und Malah.



	WA in %	WC in %
ungestört	85,4	89,3
gestört	55,7	65,1
gestört + NR	64,8	73,8

System ohne LDA



	WA in %	WC in %
ungestört	86,6	90,5
gestört	64,5	74,1
gestört + NR	60,7	77,7

System mit LDA

Abbildung 8.15: Erkennungsergebnisse zweier Systeme, eines ohne und eines mit LDA (NR bezeichnet die Geräuschreduktion nach Ephraim und Malah).

Analyse der MSE-Reduktion für verschiedene Merkmale

Bevor Erkennungsergebnisse für „echte" Autosprachdaten vorgestellt werden, soll eine Analyse mit künstlich verrauschten Aufnahmen die Eignung der Geräuschreduktion für verschiedene Merkmale abschätzen. Die Abbildungen in **Tabelle 8.1** zeigen den normierten mittleren quadratischen Fehler (NMSE, siehe Abschnitt 3.3.2, S. 45) einzelner Merkmalskoeffizienten über eine Äußerung. Die zwei Kurven zeigen jeweils den Wert pro Koeffizient, mit und ohne Geräuschreduktion. Darunter sind die Reduktionsraten für den MSE und den NMSE angegeben.

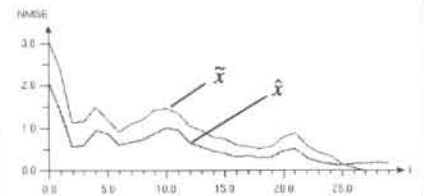
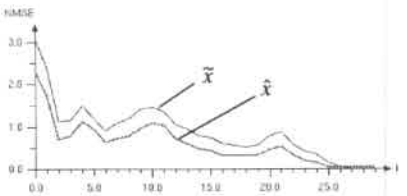
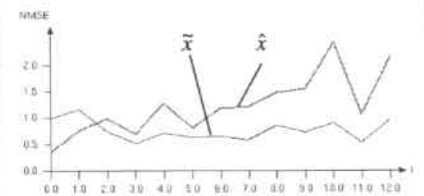
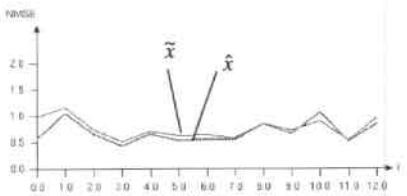
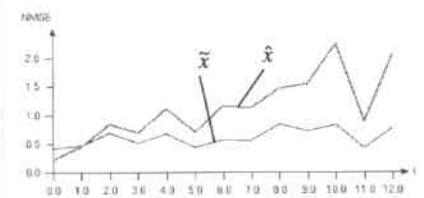
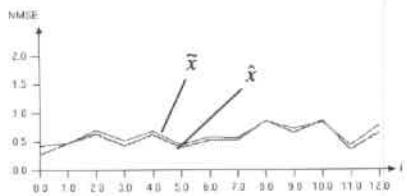
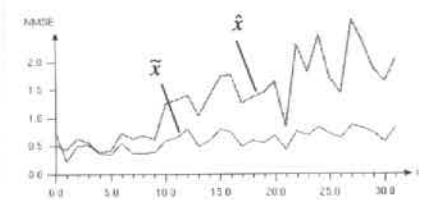
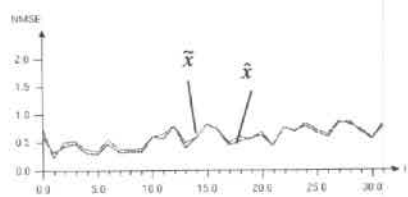
Betrachtet man zunächst die Reduktionsraten MSE_{Red}, so ergeben sich bis auf eine Ausnahme (LDA-Merkmale mit Spektraler Subtraktion) positive Werte. Das bedeutet, der euklidische Abstand zu den ungestörten Merkmalsvektoren wird im Mittel durch die Geräuschreduktion kleiner. Dieser Wert ist aber nicht für alle Merkmale aussagekräftig, denn der Fehler sagt nichts über die Unterscheidbarkeit der Klassen und auch nichts über die relative Abweichung gegenüber der „normalen" Varianz der Merkmale aus. Bei den LDA-Merkmalen sind alle Koeffizienten hingegen durch die mittlere Klassenkovarianz normiert. Hier stimmt der MSE relativ gut mit dem NMSE überein. Dieser ist im Vergleich zum MSE durch die Varianz des ungestörten Merkmals über die betrachtete Äußerung normiert.

Bei den *MEL*-Koeffizienten wird der NMSE durch beide Verfahren fast über den gesamten Spektralbereich verringert. Beim Übergang in den Cepstralbereich gilt dies nur noch für die Koeffizienten mit kleinem Index. Hierbei zeigt sich auch die Überlegenheit des Verfahrens von Ephraim und Malah gegenüber der einfachen Spektralen Subtraktion.

Die Änderung im Fehlerverhalten lässt sich dadurch erklären, dass beide Verfahren nur auf einzelnen Spektralkoeffizienten arbeiten. Benachbarte Koeffizienten werden zum Teil unterschiedlich verändert, so dass durch die Bildung der Cepstralmerkmale (Analyse des Spektralverlaufs) größere Fehler entstehen können. Die nachfolgende sprachbasierte Kanalkompensation (siehe nächster Abschnitt) gleicht gewisse Abweichungen der unteren Cepstralmerkmale weiter aus. Der Einsatz der LDA führt schließlich zu einem völlig uneinheitlichen Bild über die Koeffizienten, bei dem zwar die Spektrale Subtraktion deutlich schlechter abschneidet, aber selbst für das Verfahren nach Ephraim und Malah keine klare Aussage bezüglich des NMSE getroffen werden kann. Die klaren Vorteile, die sich für dieses Maß bei den spektralen Merkmalen ergeben haben, sind zumindest hier nicht mehr zu erkennen.

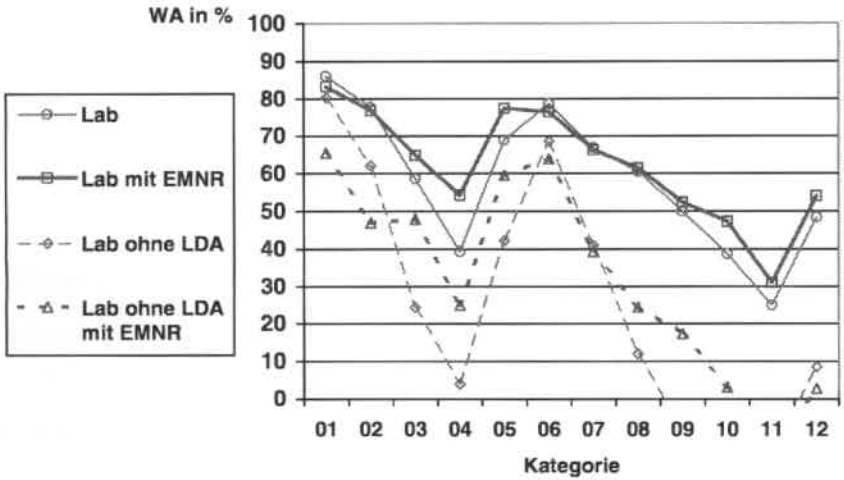
Tabelle 8.1: *MSE*-Werte eines Satzes für verschiedene Merkmale.

Die *NMSE*-Werte sind jeweils über den Koeffizienten-Index *i* aufgetragen:
dünn: gestörtes Merkmal $\tilde{x}$
dick: geräuschreduziert $\hat{x}$
Darunter sind jeweils die Reduktionsraten berechnet über alle Koeffizienten angeben.

Spektrale Subtraktion	**Ephraim & Malah**
logarithmierte *MEL*-Koeffizienten	logarithmierte *MEL*-Koeffizienten
	
MSE_{Red} = 36,1 % $NMSE_{Red}$ = 35,6 %	MSE_{Red} = 30,9 % $NMSE_{Red}$ = 31,4 %
Cepstral-Koeffizienten	Cepstral-Koeffizienten
	
MSE_{Red} = 50,7 % $NMSE_{Red}$ = -59,5 %	MSE_{Red} = 35,8 % $NMSE_{Red}$ = 9,7 %
Cepstral-Koeffizienten nach der Kanalkompensation (SCMS)	Cepstral-Koeffizienten nach der Kanalkompensation (SCMS)
	
MSE_{Red} = 25,6 $NMSE_{Red}$ = -80,4	MSE_{Red} = 27,0 $NMSE_{Red}$ = 9,2
Merkmalskoeffizienten nach der LDA	Merkmalskoeffizienten nach der LDA
	
MSE_{Red} = -101 % $NMSE_{Red}$ = -119 %	MSE_{Red} = 7,0 % $NMSE_{Red}$ = 5,6 %

Kontinuierliche Erkennung mit Autosprachdaten

Abbildung 8.16 zeigt die Ergebnisse von zwei kontinuierlichen Erkennern (mit und ohne LDA) auf realen Autosprachdaten. Daraus geht hervor, dass durch die Geräuschreduktion nach Ephraim und Malah (EMNR) für den Basiserkenner mit LDA bei stark verrauschten Kategorien durchaus Gewinne erzielt werden können. Diese fallen aber weniger deutlich aus als bei dem wesentlich schlechteren System ohne LDA.



System	aus	Blinker	Fenster	Lstark	Lmittel	0km/h	25km/h	50km/h	75km/h	100km/h	125km/h	beschl.
Basiserkenner (WA in %)	86,1	77,7	58,6	39,4	68,9	78,7	66,9	60,5	50,0	38,7	25,0	48,5
+ EMNR (WA in %)	83,3	76,8	64,9	54,2	77,4	76,5	66,3	61,5	52,4	47,3	31,1	54,0
ER relativ	-20%	-4%	15%	24%	27%	-10%	-2%	3%	5%	14%	8%	11%

Kategorie	01	02	03	04	05	06	07	08	09	10	11	12
Systeme ohne LDA	aus	Blinker	Fenster	Lstark	Lmittel	0km/h	25km/h	50km/h	75km/h	100km/h	125km/h	beschl.
Basiserkenner (WA in %)	80,3	62,1	24,4	4,0	42,3	68,5	41,0	12,0	-8,5	-13,2	-14,5	8,6
+ EMNR (WA in %)	65,5	47,0	47,9	25,0	59,5	63,9	39,3	24,4	17,4	3,1	-15,6	2,8
ER relativ	-75%	-40%	31%	22%	30%	-15%	-3%	14%	24%	14%	-1%	-6%

Abbildung 8.16: Ergebnisse der Geräuschreduktion nach Ephraim und Malah (EMNR) für System mit und ohne LDA.

8.4.4 Zusammenfassung

Obwohl sich die Geräuschreduktionsfilter bei Einzelwort-Experimenten als sehr wirksam gezeigt haben, ist die Verbesserung bei der Erkennung kontinuierlich gesprochener Anfragen noch zu gering, um eine ernsthafte Alternative zu den beiden neu trainierten Referenzsystemen (aus Abschnitt 6.2.3) zu sein. Gründe dafür mögen in der Betrachtung nur einzelner spektraler Merkmale liegen. Daneben wird die Varianz des Rauschens völlig

vernachlässigt. Von Nachteil sind auch die zahlreichen freien Parameter, die im vorliegenden Fall auf einigen Autosprachdaten eingestellt werden mussten.

Im nächsten Abschnitt wird die Kanalkompensation für additives Rauschen erweitert. Auf diese Weise werden komplette Merkmalsvektoren betrachtet. Darauf aufbauend wird im Kapitel 9 ein neues Verfahren eingeführt, das komplette Merkmalsvektoren und die Varianz des Rauschens berücksichtigt und damit sehr viel effektiver die additiven Störungen kompensieren kann.

8.5 Kanalkompensation

8.5.1 Einfache Kanalkompensation

Definition der Cepstralen Mittelwert-Subtraktion (CMS)

Laut dem Kanalmodell aus Abschnitt 8.2 schlagen sich Aussteuerung, Übertragungs- und Sprechereigenschaften als konstante Verschiebung im Cepstral- und *Log*-Spektralbereich nieder. Bei vielen Spracherkennungssystemen kommt daher die Cepstrale Mittelwert-Subtraktion (CMS) [Atal 1974, Furui 1981] zum Einsatz, deren Aufgabe es ist, den Einfluss des Kanals und die Unterschiede zwischen den Sprechern zu verringern (siehe auch Gleichung (8.8) auf Seite 131). Durch Subtraktion des Mittelwertes aller cepstralen Merkmalsvektoren einer oder mehrerer Äußerungen werden so die lautinformationstragenden Anteile des Sprachsignals gewonnen.

$$\textbf{CMS:} \qquad \begin{aligned} \mathbf{y}[k] &= \mathbf{x}[k] - \bar{\mathbf{x}} \\ &\approx \mathbf{x}[k] - \left(\tilde{\mathbf{h}}^{K} + \tilde{\mathbf{h}}^{S}\right) \end{aligned} \tag{8.17}$$

Die sonst verwendete Markierung cepstraler Merkmale mit einem hochgestellten „c" wurde hier weggelassen, da die Zusammenhänge prinzipiell auch für *Log*-Spektralmerkmale gelten. Beim Sprecherkanal $\tilde{\mathbf{h}}^{S}$ wird angenommen, dass sich mit einer gewissen Anzahl von Sprachmustern eine bestimmte sprachentypische, phonetische A-priori-Verteilung einstellt. Mit dieser Annahme ist der Mittelwert der Muster nur charakteristisch für den Sprecher und den verwendeten Aufnahmekanal, aber nicht mehr abhängig vom sprachlichen Inhalt. Je größer die Stichprobe bei konstanten Aufnahmebedingungen umso besser ist diese Voraussetzung erfüllt.

Eignung der CMS für verrauschte Aufnahmen

Abbildung 8.17 zeigt den ersten Cepstralkoeffizient aller Analyserahmen k einer Äußerung bevor (x_0^c) und nachdem ($\tilde{x}_0^c$) Fahrgeräusch zur Sprachaufnahme addiert wurde. Wie man dort sehen kann, wird der Mittelwert des resultierenden Merkmals $\bar{\tilde{x}}_0^c$ durch das additive Rauschen stark beeinflusst. Die **Abbildung 8.18** daneben zeigt die Verhältnisse beim Einsatz der CMS. In gewisser Weise sorgt sie auch für eine Kompensation der (in diesem Merkmalsraum nicht additiven) Störgeräusche, denn gerade kleine und mittlere Werte weichen weniger stark von ihrem ungestörten Pendant ab.

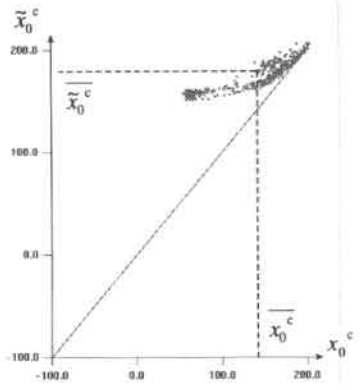


Abbildung 8.17: Auswirkung des Fahrgeräusches auf den ersten Cepstral-Koeffizienten x_0^c.

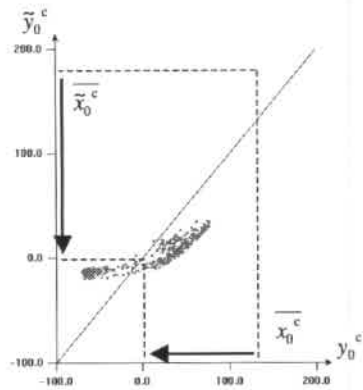


Abbildung 8.18: Wirkung von Fahrgeräusch und Mittelwertsubtraktion auf den ersten Cepstral-Koeffizient x_0^c.

Sprachbasierte CMS

Eine Schwäche der einfachen Cepstralen Mittelwertsubtraktion ist ihre starke Abhängigkeit von der Dauer der Sprachpausen und dem Störgeräusch. Abweichungen der Kanalschätzung entstehen durch unterschiedliche Anteile von Sprachpausen bei der Mittelwertbildung. Je größer deren Menge umso kleiner wird in der Regel der geschätzte Mittelwert. Ist das Signal durch additives Rauschen gestört, führt dies zu einer Vergrößerung des Schätzwertes. Um dieser Abhängigkeit zu entkommen, kann man sich bei der Auswahl der zu mittelnden Vektoren auf die Sprachmuster beschränken. Die Klassifizierung muss mithilfe eines Sprachdetektors durchgeführt werden. Dieser liefert entweder eine binäre Entscheidung oder ein kontinuierliches Maß w[k] zwischen 0 und 1. Damit ergibt sich die Definition der sprachbasierten CMS wie folgt:

$$\textbf{SCMS:} \qquad \mathbf{y}[k] = \mathbf{x}[k] - \overline{\mathbf{x}}^{\text{Sprache}} \quad \text{mit} \quad \overline{\mathbf{x}}^{\text{Sprache}} = \frac{\sum_k \mathrm{w}[k] \cdot \mathbf{x}[k]}{\sum_k \mathrm{w}[k]}. \tag{8.18}$$

Da sich die Energie eines Merkmalsvektors in x_0^c widerspiegelt, kann man an diesem Merkmal zumindest ungefähr Sprach- und Pausenrahmen unterscheiden. Sprachrahmen zeichnen sich meist durch größere Werte des ungestörten Merkmals x_0^c aus. Wie **Abbildung 8.19** und **Abbildung 8.20** zeigen, ergibt sich im Vergleich zur einfachen CMS nun eine bessere Übereinstimmung von gestörten und ungestörten Signal für die Sprachanteile, während die Pausen wie im Falle ohne CMS stark abweichen.

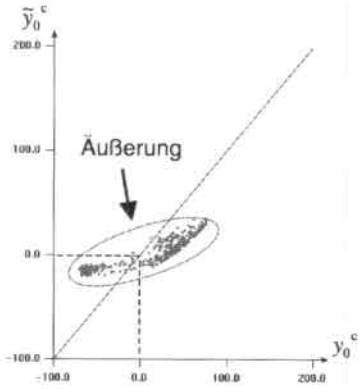


Abbildung 8.19: Geräuscheinfluss auf Cepstral-Koeffizient x_0^c mit einfacher CMS

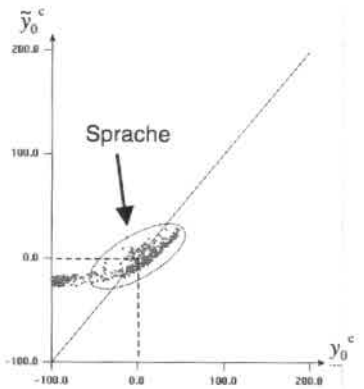


Abbildung 8.20: Geräuscheinfluss auf Cepstral-Koeffizienten x_0^c mit sprachbasierter CMS

Vergleich von CMS und SCMS für Labordaten

Tabelle 8.2 zeigt Erkennungsraten zum Vergleich der bisher vorgestellten Kanalkompensationsverfahren für rauscharme Aufnahmen (Kategorie 01) auf den beiden zur Sprachdatensammlung im Auto benutzten Mikrofonen. „Sennheiser" ist ein Nahbesprechungsmikrofon, das auch für die Aufnahme der GSST-Trainingsdaten verwendet wurde. „AKG" ein in den Testwagen installiertes Raummikrofon. Bei Aufnahmen mit dem „Sennheiser" erweist sich die SCMS als überlegen. Die Wortakkuratheit WA verbessert sich von 85,6 % auf 86,9 %, dies entspricht einer relativen Fehlerreduktion von 9 %. Hier kommt wohl der Vorteil zum tragen, unabhängiger vom Anteil der Sprachpausen zu sein.

Das System ganz ohne CMS erweist sich für die Sennheiser-Aufnahmen überraschend gut, bricht aber mit dem AKG deutlich ein. Anders dagegen die Systeme mit Kanalkompensation. Beide liefern etwa die gleiche Anzahl korrekter Wörter (WC). Da bei der SCMS die Merkmalsvektoren der Pausenregionen kaum ausgeglichen werden, erhöht sich durch das leichte Rauschen des AKGs die Zahl der Einfügefehler, was zu einem Absinken der Wortakkuratheit führt. Dennoch ist die WC-Rate immer noch höher als bei der CMS.

Tabelle 8.2: Ergebnisse der Kanalkompensation für zwei verschiedene Mikrofone.

Mikrofon: / Kompensation:	Sennheiser		AKG	
	WA in %	WC in %	WA in %	WC in %
Keine	85,3	89,8	71,8	77,1
CMS	85,6	89,9	86,4	89,8
sprachbasierte CMS (SCMS)	86,9	91,1	**86,1 *)**	**91,4 *)**

***)** Vergleichssystem (Basiserkenner) aller sonstiger Untersuchungen

8.5.2 Kanalkompensation für verrauschte Daten

Verwendung zweier Mittelwerte

Um die guten Eigenschaften der sprachbasierten CMS auch für verrauschte Daten nutzbar zu machen, wird nun ein Weg beschrieben, die besonders gestörten Pausenmuster ebenfalls zu kompensieren. Unter Ignorierung der additiven Eigenschaft der Störgeräusche im Spektralbereich wird ein weiterer Ausgleichsvektor für die cepstralen Merkmalsvektoren der Pausen eingeführt. Dieser Vektor wird durch Mittelwertbildung auf den als Pause klassifizierten Merkmalsvektoren geschätzt und auch nur von diesen subtrahiert. Diese 2CMS (engl.: *2-level cepstral mean subtraction*) mit binärer Funktion $w[k]$ ergibt sich somit zu

2CMS:
$$\mathbf{y}[k] = \mathbf{x}[k] - w[k] \cdot \bar{\mathbf{x}}^{\text{Sprache}} - (1 - w[k]) \cdot \bar{\mathbf{x}}^{\text{Pause}} \tag{8.19}$$

$$\text{mit } \bar{\mathbf{x}}^{\text{Sprache}} = \frac{\sum_k w[k] \cdot \mathbf{x}[k]}{\sum_k w[k]} \quad \text{und} \quad \bar{\mathbf{x}}^{\text{Pause}} = \frac{\sum_k (1 - w[k]) \cdot \mathbf{x}[k]}{\sum_k (1 - w[k])}.$$

Mit diesem Verfahren werden bei [Gupta *et al.* 1996] für einen Ziffernerkenner in der Autoumgebung relative Fehlerreduktionen zwischen 7 und 20 % erreicht. Andererseits schneidet es bei [Wittmann *et al.* 1993] im Vergleich zur SCMS schlechter ab. Auch wenn also für manche Anwendungen Verbesserungen erzielt werden, ergibt sich ein schwerwiegender Seiteneffekt: Sprach- und Pausenmuster werden durch die jeweilige Mittelwertsubtraktion zum Ursprung verschoben und überlagern sich dort. Ohne ein weiteres Energiemerkmal häufen sich dann bei einem Erkenner für kontinuierliche Sprache die Verwechslungen von Pausen mit Sprachlauten (Einfügefehler!). Außerdem ist die Merkmalsextraktion sehr stark von der richtigen Zuordnung des verwendeten Sprachdetektors abhängig.

Daher wird dieser Ansatz nun um zwei entscheidende Punkte erweitert: Die Subtraktion wird durch eine Normierung ersetzt und als Gewichtungsfunktion $w[k]$ ein kontinuierliches, energiebasiertes Maß verwendet. Die beiden Mittelwerte werden nicht wie oben zum Nullpunkt verschoben, sondern zum durchschnittlichen Wert aller Sprach- bzw. Pausenmuster der Trainingsmenge. Deshalb wird als Ausgleichsvektor die Differenz (Delta) vom aktuellen zum durchschnittlichen Mittelwert aller Trainingsäußerungen verwendet. Die Definition dieser zweifachen Delta-CMS (2DCMS) [Westphal 1997] lautet

2DCMS:
$$\begin{aligned}\mathbf{y}[k] &= \mathbf{x}[k] - w[k] \cdot \left(\bar{\mathbf{x}}^{\text{Sprache}} - \boldsymbol{\mu}_{\text{Sprache}} \right) - (1 - w[k]) \cdot \left(\bar{\mathbf{x}}^{\text{Pause}} - \boldsymbol{\mu}_{\text{Pause}} \right) \\ &= \mathbf{x}[k] - w[k] \cdot \left(\Delta\bar{\mathbf{x}}^{\text{Sprache}} \right) - (1 - w[k]) \cdot \left(\Delta\bar{\mathbf{x}}^{\text{Pause}} \right)\end{aligned} \tag{8.20}$$

$$\text{mit } \bar{\mathbf{x}}^{\text{Sprache}} = \frac{\sum_k w[k] \cdot \mathbf{x}[k]}{\sum_k w[k]} \quad \text{und} \quad \bar{\mathbf{x}}^{\text{Pause}} = \frac{\sum_k (1 - w[k]) \cdot \mathbf{x}[k]}{\sum_k (1 - w[k])}$$

und N Trainingsäußerungen
mit den Merkmalsvektoren $\mathbf{x}_u[k]$ für die Äußerung u:

$$\mu_{\text{Sprache}} = \frac{1}{N}\sum_{u=1}^{N} \mathbf{x}_u^{-\text{Sprache}} \quad \text{und} \quad \mu_{\text{Pause}} = \frac{1}{N}\sum_{u=1}^{N} \mathbf{x}_u^{-\text{Pause}}$$

Abbildung 8.21 und **Abbildung 8.22** zeigen jeweils die Auswirkungen einer diskreten und kontinuierlichen Realisierung der 2DCMS. Bei der diskreten Umsetzung ist ein deutlicher Bruch zwischen den als Pause bzw. Sprache klassifizierten Merkmalen zu sehen. Daher sollte der verwendete Klassifikator eine sehr kleine Fehlerrate haben. Im kontinuierlichen Fall ergibt sich dagegen ein allmählicher Übergang. Da die Auswirkung des additiven Rauschens sich unabhängig von der Klassenzugehörigkeit stärker auf energieschwache Merkmalsvektoren auswirkt, reicht ein einfaches Energiemaß für diese Zwecke aus. Die korrekte gleichmäßige Auswahl zumindest einer Teilmenge der Sprachmuster durch die Funktion w[k] ist hingegen für eine gute Schätzung des Kanals erforderlich. In den gezeigten Beispielen fällt die Schätzung des Sprachmittelwerts im Vergleich zu den ungestörten Daten bedingt durch das addierte Rauschen etwas zu groß aus. Würde man sich allerdings auf einen kleineren Teil (energiereicher) Beispiele stützen, die weniger gestört sind, ständen bei sehr kurzen Äußerungen nur wenige und daher stark äußerungsabhängige Daten zur Verfügung.

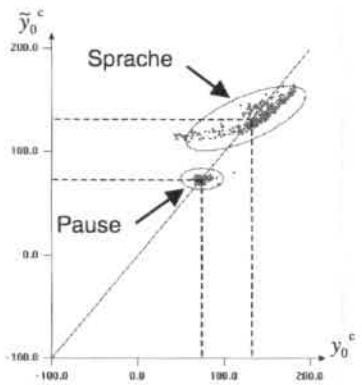


Abbildung 8.21: Geräuscheinfluss auf Cepstral-Koeffizienten $x_0^{\,c}$ mit diskreter 2DCMS.

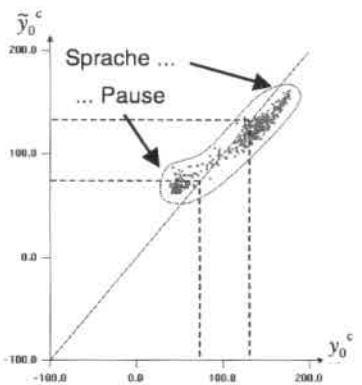


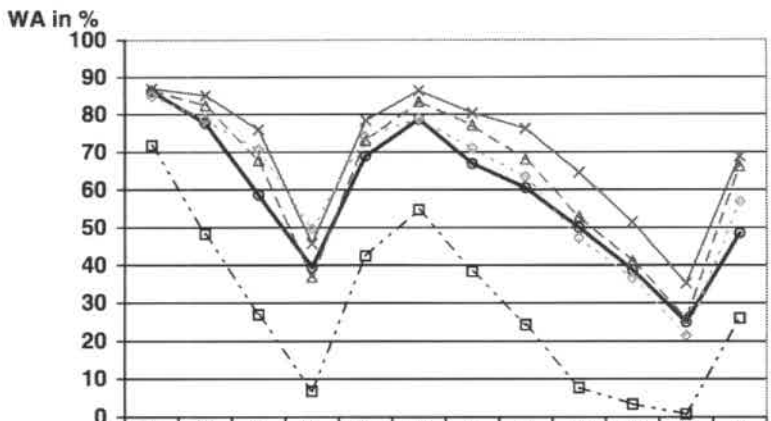
Abbildung 8.22: Geräuscheinfluss auf Cepstral-Koeffizienten $x_0^{\,c}$ mit kontinuierlicher 2DCMS.

Versuchsergebnisse

Ergebnisse für die vorgestellten Kanalkompensationsverfahren sind in **Abbildung 8.23** dargestellt. Bei allen Versuchen war aufgrund der unterschiedlichen Merkmale ein Neutraining erforderlich. Wie auch schon anhand der zuvor besprochenen Beispiele zu vermuten war, schneidet die kontinuierliche 2DCMS bei verrauschten Daten am besten ab. Während die sprachbasierte CMS (SCMS) bei Aufnahmen mit dem Sennheiser-Mikrofon noch bessere Ergebnisse lieferte, unterliegt sie hier der einfachen CMS in den meisten Kategorien.

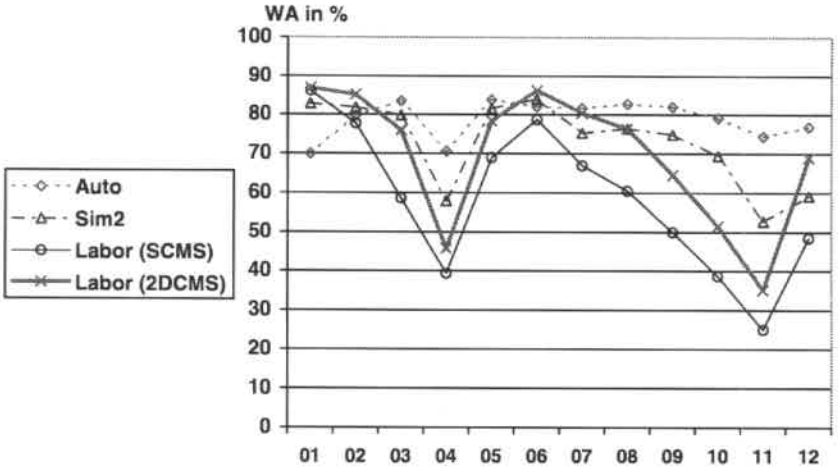
Neben der einfachen CMS ist noch eine weitere Variante (CMSvar) aufgeführt, bei der nicht nur der Mittelwert subtrahiert wurde, sondern auch für jede Äußerung die Varianz normiert wurde. Dieses Vorgehen wird in einigen Veröffentlichungen (z.B. [Shieh und Chang 1999]) als der CMS überlegen beschrieben. Für das hier betrachtete Erkennungssystem bringt es aber keine Vorteile, wahrscheinlich weil damit bei anderen Erkennern hauptsächlich unterschiedlich dimensionierte Merkmale ausgeglichen werden. Dies wird aber im vorliegenden Fall bereits durch die Verwendung der LDA-Transformation bewerkstelligt.

Alle kompensierten Systeme schneiden deutlich besser ab als ein System ohne Kanalkompensation (noCMS). Waren bei der Verwendung des Sennheiser-Mikrofons unter Laborbedingungen kaum Einbußen zu verzeichnen, reduziert sich hier die Wortakkuratheit für das AKG bis auf beinahe 0 % (Kategorie 11, Tempo 125 km/h).



	01	02	03	04	05	06	07	08	09	10	11	12
Besonderheit	aus	Blinker	Fenster	Lstark	Lmittel	0km/h	25km/h	50km/h	75km/h	100km/	125km/	beschl.
Labor (SCMS)	86,1	77,7	58,6	39,4	68,9	78,7	66,9	60,5	50	38,7	25	48,5
Labor (2DCMS)	86,9	85,1	75,9	45,8	78,4	86,3	80,4	76,3	64,5	51,5	35,2	68,9
Labor (CMS)	86,4	82,4	67,8	36,8	73,1	83,4	77,1	68,1	52,9	41,1	26	66,2
Labor (CMSvar)	84,9	78,7	70,6	49,7	74,2	78,9	71	63,4	47,2	36,5	21,6	56,9
Labor (noCMS)	71,8	48,3	27	6,8	42,5	54,8	38,4	24,3	7,7	3,4	0,8	26

Abbildung 8.23: Ergebnisse der Kanalkompensationsverfahren mit dem im Auto installierten Mikrofon (AKG).



Kategorie	01	02	03	04	05	06	07	08	09	10	11	12
System	aus	Blinker	Fenster	Lüfter	Lüftrel	0kmh	25kmh	50kmh	75kmh	100kmh	125kmh	beschl.
Basiserkenner (WA in %)	86,1	77,7	58,6	39,4	68,9	78,7	66,9	60,5	50,0	38,7	25,0	48,5
mit 2DCMS (WA in %)	86,9	85,1	75,9	45,8	78,4	86,3	80,4	76,3	64,5	51,5	35,2	68,9
ER relativ	6%	33%	42%	11%	31%	36%	41%	40%	29%	21%	14%	40%

Abbildung 8.24: Vergleich des Laborerkenners mit 2DCMS-Kanalkompensation mit den Bezugssystemen.

In **Abbildung 8.24** wird das unter realen Bedingungen beste Kanalkompensationsverfahren 2DCMS mit den Referenzsystemen aus Abschnitt 6.2.3 verglichen (Basiserkenner „Labor" und die mit gestörten Daten trainierten Systeme „Sim2" und „Auto"). Diese wurden mit sprachbasierter bzw. einfacher CMS trainiert und getestet. Durch die verbesserte Kanalkompensation werden gegenüber dem Basiserkenner Fehlerreduktionen bis über 40 % erzielt. Auf leicht gestörten Kategorien werden auch die Ergebnisse der anderen Referenzsysteme übertroffen. Für stark gestörte Kategorien kann jedoch dieses Verfahren die Erkennungsraten des Laborerkenners nicht entscheidend verbessern. Motiviert durch die dennoch beachtlichen Ergebnisse wird in Kapitel 9 (Modellbasierte Umgebungskompensation) das Konzept der 2DCMS nochmals aufgenommen.

8.5.3 Zusammenfassung

Zusammenfassend lässt sich sagen, dass die Verfahren basierend auf den cepstralen Mittelwerten einem Spracherkenner zum Teil auch eine gewisse Robustheit gegenüber additivem Rauschen verleihen. Jedoch wird durch diese andersartige Störung der eigentliche Zweck erschwert, nämlich die Schätzung und Kompensation von wechselnden (für eine Äußerung aber konstante) Kanaleigenschaften.

Wird der Mittelwert nur auf Sprachmustern bestimmt (SCMS) bzw. auf Mustern, die hohe Energie besitzen, erhält man eine gute Schätzung für den Kanal, die nicht durch die Länge der Pausen oder das darin enthaltene Störgeräusch beeinflusst wird.

Mit mehr als einem Mittelwert kann man auch eine Kompensation von additivem Störgeräusch erreichen. Dabei werden nicht nur einzelne Koeffizienten unabhängig voneinander kompensiert, sondern der Merkmalsvektor als Ganzes. Die 2DCMS führte daher zu besseren Ergebnissen als die einkanaligen Geräuschreduktionsfilter.

9 Modellbasierte Umgebungskompensation

Wir brauchen keine Theorie,
die die Nase rümpft, wenn sie der Praxis begegnet.
- Prof. Dr. h.c. Manfred Rommel -

In diesem Kapitel wird basierend auf den bisher beschriebenen Erkenntnissen ein neues Verfahren (MAM) zur Kompensation der Umgebungseinflüsse vorgestellt. Der erste Abschnitt 9.1 beschreibt die Überlegungen, die zu diesem Verfahren geführt haben, und klärt die grundsätzliche Vorgehensweise. Im zweiten Abschnitt 9.2 wird anhand eines Umgebungsmodells, das die bisher betrachteten Störungen gemeinsam beschreibt, die praktische Umsetzung und deren Umgang mit den Störeinflüssen vorgestellt. Danach folgt in Abschnitt 9.3 eine experimentelle Auswertung, bei dem das Verfahren seine praktische Eignung unter Beweis stellt. Abschnitt 9.4 zieht einen Vergleich des Verfahrens mit anderen bisher vorgestellten Ansätzen und beendet das Kapitel mit einer abschließenden Bewertung.

9.1 Modellkombinationsbasierte Akustische Transformation (MAM)

Die bisher nur einzeln angegangenen Störeinflüsse *additives Rauschen*, *Kanal* und *Sprecher* sollen bei einer allgemeinen Umgebungskompensation berücksichtigt werden. Folgende Überlegungen, meist Nachteile der bisher diskutierten Verfahren, spielten beim Entwurf eine Rolle:

- Die **Filtermethoden zur Reduktion des additiven Rauschens** (Abschnitt 8.4, S. 137) arbeiten nur eindimensional. Ist ein Koeffizient durch das Rauschen stark überlagert, ist es durch die eindimensionale Betrachtung praktisch nicht mehr möglich einen guten Schätzwert zu finden (*Missing Feature*, siehe S. 133). Das Rauschen wird bei den Filtermethoden außerdem nur über einen Mittelwert modelliert, die Varianz hingegen wird nicht berücksichtigt. Durch lokale Ausgleichsmaßnahmen, wie zum Beispiel die Beschränkung durch ein Restrauschen, wird die ursprüngliche Struktur der spektralen Merkmalsvektoren zerstört. Je nach Merkmalsraum wird die Abweichung zum ungestörten Merkmalsvektor dadurch größer (vergleiche **Tabelle 8.1**, S. 146). Kompliziertere und damit meist bessere Filterverfahren weisen eine Fülle von freien Parametern auf, die je nach Anwendungsumgebung experimentell eingestellt werden müssen.

- Das additive Rauschen stört die Kanalschätzung bei der **Kanalkompensation**. Wird der Kanal wie bei der sprachbasierten CMS besser geschätzt, werden im Vergleich mit der einfachen CMS die Abweichungen zu den ungestörten Merkmalen in den Pausenregionen größer. Die Berücksichtigung dieser Umstände ermöglicht das gute Abschneiden der 2DCMS (Abschnitt 8.5.2, S. 151). Im Gegensatz zu den Filterverfahren wird hier jeweils der gesamte Merkmalsvektor bei der Kompensation berücksichtigt. Die Umsetzung dieses Verfahrens beruht allerdings nur auf einem sehr groben Umgebungsmodell (s.u.).
- Die **Sprechernormierung** (Kapitel 7) mit den modellbasierten Ansätzen bringt die besten Ergebnisse, ist aber sehr rechenaufwändig. Beruht die Bestimmung des VTLN-Verzerrungsparameters auf einer Einteilung der Muster in phonetische Klassen, ist zunächst ein eigener Erkennungslauf nötig, um diese Zuordnung zu finden. Bei additivem Störgeräusch ist die Zuordnung jedoch unzuverlässig und das akustische Modell für ungestörte Sprache, das zur Bestimmung des Verzerrungsfaktors verwendet wird, passt außerdem nicht mehr zu den gestörten Mustern.
- **Adaptionsverfahren** (Kapitel 6) können sehr effektiv sein (siehe OSTD, S. 99), da sie eine Beziehung zwischen veränderten Merkmalen und dem akustischen Modell des Erkenners herstellen. Mit Beispieldaten aus der neuen Umgebung werden eine oder mehrere Abbildungen gefunden, um dann alle Dichten des akustischen Modells des Erkenners zu verändern. Für eine gute Anpassung werden jedoch sehr viele Adaptionsdaten benötigt und der Rechenaufwand ist bei großen Erkennermodellen sehr hoch.

Um einerseits die aufgeführten Nachteile zu beseitigen und andererseits Vorteile aus den jeweiligen Stärken zu ziehen, wurde nach einer geeigneten Kombination der unterschiedlichen Ansätze gesucht. Als Ausgangspunkt für eine Kompensation des additiven Rauschens wird im nächsten Abschnitt 9.1.1 das Prinzip der bereits vorgestellten zweistufigen Kanalkompensation 2DCMS und der dort benutzten Transformation betrachtet. Der folgende Abschnitt 9.1.2 beschreibt eine erweiterte Version für M implizite Klassen. Die in Abschnitt 9.1.3 vorgeschlagene Lösung, wie diese Transformation letztendlich zur Störkompensation eingesetzt werden kann, bildet das Grundprinzip eines neuen Verfahrens, das dann im Weiteren theoretisch und praktisch untersucht wird.

9.1.1 Prinzip der erweiterten Kanalkompensation

Zunächst wird dargestellt, welches Umgebungsmodell der 2DCMS zugrunde liegt und danach durch welches Prinzip die Störungen ausgeglichen werden. Da das Cepstrum aus einer linearen Transformation des logarithmierten Spektrums hervorgeht, sind diese beiden Merkmalsräume in Hinsicht auf die Mittelwert-Subtraktionsverfahren gleichwertig. Hier wird in Übereinstimmung mit den später verwendeten Merkmalen daher das *Log*-Spektrum verwendet (angedeutet durch hochgestelltes „l").

Mit dem Kanalmodell aus Abschnitt 8.2 ergibt sich eine Beziehung zwischen dem informationstragenden Sprachanteil **s** und dem Kanal **h** mit additivem Rauschen **n**. Im Spektral-Bereich (angedeutet durch ein hochgestelltes „s") gilt danach für die Koeffizienten:

$$x_i^s = h_i^s \cdot s_i^s + n_i^s \tag{9.1}$$

Durch eine Unterteilung in die beiden Klassen *Sprache* und *Pause* kann man für die Merkmale im *Log*-Spektralbereich eine sehr grobe Näherung folgender Art annehmen:

$$\text{Sprache:} \qquad x_i^l \approx h_i^l + s_i^l \qquad \text{denn} \quad h_i^s \cdot s_i^s >> n_i^s \tag{9.2}$$

$$\text{Pause:} \qquad x_i^l = n_i^l \qquad \text{denn} \quad h_i^s \cdot s_i^s = 0 \tag{9.3}$$

Bei der 2DCMS werden die ungestörten Muster $\mathbf{x}^l$ durch ein sehr einfaches akustisches Modell λ mit diesen zwei Klassen beschrieben. Es besteht aus den Mittelwerten für Sprachemuster μ_h und Pausenmuster μ_n, die aus den Trainingsdaten gewonnenen wurden.

Modell ungestörter Sprache: $\lambda = \{\mu_h, \mu_n\}$

Der Mittelwert μ_h entspricht dabei dem Standardkanal $\mathbf{h}^l$, der sich aus dem eigentlichen Übertragungskanal und dem mittleren *Log*-Spektrum der Sprache (Sprecherkanal, siehe S. 131) ergibt. Der Mittelwert der Pause μ_n entspricht dem mittleren logarithmierten Restrauschen $\mathbf{n}^l$.

Im Anwendungsfall (also während der Erkennung) wird aus den dort vorgefundenen gestörten Mustern $\tilde{\mathbf{x}}^l[k]$ ein neues Modell bestimmt.

Modell gestörter Sprache: $\tilde{\lambda} = \{\tilde{\mu}_h, \tilde{\mu}_n\}$

Dazu müssen zunächst die Muster den Klassen zugeordnet werden, beispielsweise durch Näherungen der A-posteriori-Wahrscheinlichkeiten $P(\text{Sprache}|\tilde{\mathbf{x}}^l[k])$ und $P(\text{Pause}|\tilde{\mathbf{x}}^l[k])$. Über diese Zuordnung gewichtet können dann aus den Mustern die Modellparameter (also die Mittelwerte) berechnet werden.

Über die schließlich durchgeführte gewichtete Subtraktion werden aus den gestörten Daten die ungestörten geschätzt. Mit der neuen, hier verwendeten Notation ergibt sich die Schätzformel für die **2DCMS** aus Abschnitt 8.5.2 zu

$$\hat{\mathbf{x}}[k] = \tilde{\mathbf{x}}[k] - P(\text{Sprache}|\tilde{\mathbf{x}}^l[k]) \cdot (\tilde{\mu}_h - \mu_h) - P(\text{Pause}|\tilde{\mathbf{x}}^l[k]) \cdot (\tilde{\mu}_n - \mu_n). \tag{9.4}$$

Es handelt sich hierbei um eine *MMSE*-Schätzung (siehe Abschnitt 3.3.2), die im folgenden Abschnitt für Modelle mit M Klassen hergeleitet wird. Die 2DCMS kann man als eine einfache Variante mit $M = 2$ betrachten.

9.1.2 Akustische Transformation

Die vorliegenden Merkmalsvektoren $\tilde{\mathbf{x}}[k]$ stellen gestörte Versionen von $\mathbf{x}[k]$ dar. Den Zusammenhang kann man über einen Verschiebungsvektor $\boldsymbol{\Delta}$ beschreiben, der allerdings im Gegensatz zur einfachen CMS nicht konstant angenommen wird, sondern vom Merkmalsvektor $\mathbf{x}[k]$ abhängt.

$$\tilde{\mathbf{x}}[k] = \mathbf{x}[k] + \boldsymbol{\Delta}(\mathbf{x}[k]) \tag{9.5}$$

Über den Verschiebungsvektor werden keine weiteren Annahmen gemacht. Es liegen jedoch prototypische Vektoren $\boldsymbol{\mu}_m$ und $\tilde{\boldsymbol{\mu}}_m$ vor. Diese stellen generische Modelle λ (allgemeines Modell für ungestörte Sprache) und $\tilde{\lambda}$ (für gestörte Sprache) dar und bestehen jeweils aus den korrespondierenden Klassen $m = 1 \ldots M$.

Modelle mit Prototypen: $\lambda = \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_M\}, \quad \tilde{\lambda} = \{\tilde{\boldsymbol{\mu}}_1, \ldots, \tilde{\boldsymbol{\mu}}_M\}$

Es wird angenommen, dass für die obigen Muster und die sie repräsentierenden Prototypen die gleiche Beziehung gilt:

$$\tilde{\boldsymbol{\mu}}_m = \boldsymbol{\mu}_m + \boldsymbol{\Delta}(\boldsymbol{\mu}_m). \tag{9.6}$$

Der MMSE-Schätzwert $\hat{\mathbf{x}}[k]$ für das ungestörte Muster $\mathbf{x}[k]$ ergibt sich als Erwartungswert der A-posteriori-Dichte (siehe Gleichung (3.22), S. 44). Das Integral über den Merkmalsraum der Muster $\mathbf{x}$ wird hier allerdings näherungsweise durch eine Summation über die Prototypen von λ ersetzt.

$$\begin{aligned}
\hat{\mathbf{x}} &= E\{\mathbf{x} \mid \tilde{\mathbf{x}}\} = \int_{\mathbf{x}} \mathbf{x} \cdot p(\mathbf{x} \mid \tilde{\mathbf{x}}) \mathbf{dx} \\
&= \tilde{\mathbf{x}} - \int_{\mathbf{x}} \boldsymbol{\Delta}(\mathbf{x}) \cdot p(\mathbf{x} \mid \tilde{\mathbf{x}}) \mathbf{dx} \\
&\approx \tilde{\mathbf{x}} - \sum_{m=1}^{M} \boldsymbol{\Delta}(\boldsymbol{\mu}_m) \cdot P(m \mid \tilde{\mathbf{x}})
\end{aligned} \tag{9.7}$$

Aus Gleichung (9.6) ergeben sich die Verschiebungsvektoren $\boldsymbol{\Delta}(\boldsymbol{\mu}_m) = \tilde{\boldsymbol{\mu}}_m - \boldsymbol{\mu}_m$, die gewichtet mit der bedingten Klassenwahrscheinlichkeit $P(m \mid \tilde{\mathbf{x}}[k])$ einen Ausgleichsvektor $\boldsymbol{\Delta}[k]$ bilden.

$$\boldsymbol{\Delta}[k] = \sum_{m=1}^{M} (\tilde{\boldsymbol{\mu}}_m - \boldsymbol{\mu}_m) \cdot P(m \mid \tilde{\mathbf{x}}[k]) \tag{9.8}$$

Die Mittelwerte zur Bestimmung des Verschiebungsvektors sind durch die Modelle gegeben. Es bleibt die Frage wie $P(m \mid \tilde{\mathbf{x}}[k])$ bestimmt werden kann. Bei der 2DCMS reichte bei zwei Klassen dazu ein einfacher Sprach-Pause-Detektor oder ein energiebasiertes Maß. Bei mehr Klassen kann man, wie für die akustischen Modelle eines Spracherkenners, eine Mischverteilung aus Gaußdichten verwenden. Die Klassen entsprechen den einzelnen Gaußdichten und müssen hierbei nicht explizit vorgegeben sein, sondern ergeben sich datengetrieben durch ein iteratives EM-Training. Das Modell λ erweitert sich also um die Kovarianzmatrizen $\mathbf{C}_m$ und die Mixturgewichte $\hat{P}_m(m)$.

Modell ungestörter Sprache: $\lambda = \{\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_M,\mathbf{C}_1,\ldots,\mathbf{C}_M,\hat{P}_m(1),\ldots,\hat{P}_m(M)\}$

Für das gestörte Modell $\tilde{\lambda}$ ändern sich dem gegenüber lediglich die Mittelwerte und Kovarianzmatrizen (die Parameter der Gaußdichten), nicht aber die Mixturgewichte $\hat{P}_m(m)$, die mit denen von λ übereinstimmen.

Über die Auswertung einer Gaußdichte von $\tilde{\lambda}$ erfolgt die Berechnung von $\hat{p}(\tilde{\mathbf{x}}[k]|m)$. Die gesuchte Wahrscheinlichkeit lässt sich damit folgendermaßen berechnen:

$$\hat{P}(m|\tilde{\mathbf{x}}[k]) = \frac{\hat{P}_m(m)\cdot\hat{p}(\tilde{\mathbf{x}}[k]|m)}{\hat{p}(\tilde{\mathbf{x}}[k])} = \frac{\hat{P}_m(m)\cdot\hat{p}(\tilde{\mathbf{x}}[k]|m)}{\sum_{n=1}^{M}\hat{P}_m(n)\cdot\hat{p}(\tilde{\mathbf{x}}[k]|n)} \tag{9.9}$$

*Diese Art der Schätzung des ungestörten Merkmalsvektors wird hier als „**Akustische Transformation**" **AM** (engl. acoustic mapping) bezeichnet und ist wie folgt definiert:*

$$\hat{\mathbf{x}}[k] = AM(\tilde{\mathbf{x}}[k];\lambda,\tilde{\lambda}) = \tilde{\mathbf{x}}[k] + \frac{\sum_{m=1}^{M}\hat{P}_m(m)\cdot N(\tilde{\mathbf{x}}[k];\tilde{\boldsymbol{\mu}}_m,\tilde{\mathbf{C}}_m)\cdot(\boldsymbol{\mu}_m-\tilde{\boldsymbol{\mu}}_m)}{\sum_{m=1}^{M}\hat{P}_m(m)\cdot N(\tilde{\mathbf{x}}[k];\tilde{\boldsymbol{\mu}}_m,\tilde{\mathbf{C}}_m)} \tag{9.10}$$

$$\lambda = \{\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_M,\mathbf{C}_1,\ldots,\mathbf{C}_M,\hat{P}_m(1),\ldots,\hat{P}_m(M)\}$$

$$\tilde{\lambda} = \{\tilde{\boldsymbol{\mu}}_1,\ldots,\tilde{\boldsymbol{\mu}}_M,\tilde{\mathbf{C}}_1,\ldots,\tilde{\mathbf{C}}_M,\hat{P}_m(1),\ldots,\hat{P}_m(M)\}$$

λ und $\tilde{\lambda}$ sind dabei zwei korrespondierende Modelle für ungestörte und gestörte Sprache und als Gauß'sche Mischverteilung realisiert.

Das Prinzip ist graphisch in **Abbildung 9.1** veranschaulicht. Nachdem die Zuordnungen $\hat{P}(m|\tilde{\mathbf{x}}[k])$ des gestörten Merkmalsvektors $\tilde{\mathbf{x}}[k]$ zu den Gaußdichten des gestörten Modells $\tilde{\lambda}$ berechnet sind, ergibt sich ein Korrekturvektor zur Schätzung von $\hat{\mathbf{x}}[k]$. Der Korrekturvektor berechnet sich als gewichtete Summe der Mittelwert-Differenzvektoren $\boldsymbol{\mu}_m - \tilde{\boldsymbol{\mu}}_m$. Wie man die beiden korrespondierenden Modelle erhält, wird der nächste Abschnitt klären.

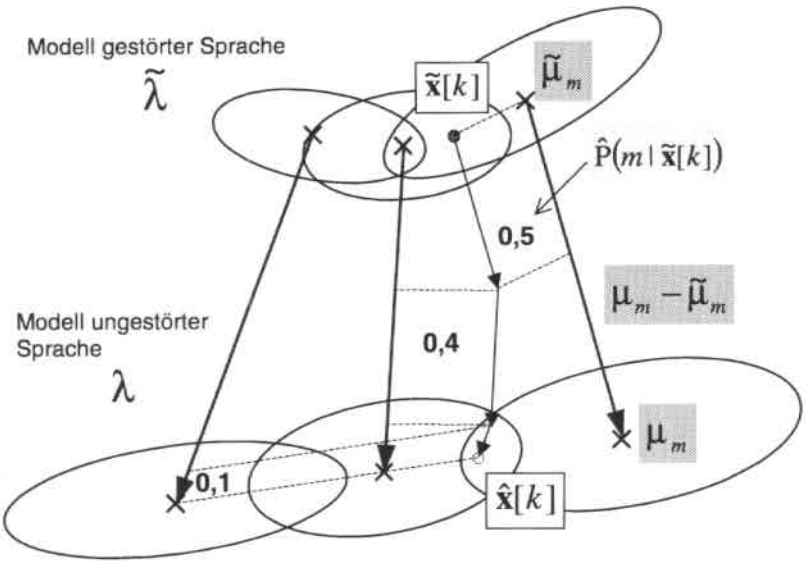


Abbildung 9.1: Veranschaulichung der Akustischen Transformation.

9.1.3 Motivation und Grundprinzip der MAM

Die Verwendung der oben hergeleiteten Akustischen Transformation (AM) setzt zwei Modelle mit korrespondierenden Klassen voraus. In [Moreno *et al.* 1995] wird dazu eine Möglichkeit mit „Stereodaten" (also simultanen Aufnahmen über zwei Kanäle) beschrieben. Dabei wird für den Standardkanal (Zielkanal) neben dem akustischen Modell des Spracherkenners auch ein generisches Hilfsmodell trainiert. Die Zuordnungen der Muster, die sich daraus für die einzelnen Dichten m ergeben, werden auch für die Daten des zweiten Kanals verwendet, für die so ein korrespondierendes Hilfsmodell erstellt werden kann. In der Anwendung werden Daten, die mit dem zweiten Kanal aufgenommen wurden, zur Schätzung der entsprechenden Daten des Standardkanals verwendet.

Für Sprachdaten, die in der Erkennungsphase über einen anderen, unbekannten Kanal aufgenommen werden, bedient man sich in [Giuliani 1999] der MLLR (Abschnitt 6.3) zur Erzeugung eines korrespondierenden Modells. Da die lineare Anpassung über einen Vergleich der andersartigen Muster mit dem generischen Sprachmodell der Trainingsdaten erfolgt, dürfen die Unterschiede jedoch nicht groß sein. Die Zuordnung erfolgt ja nur aufgrund einzelner Muster ohne die Unterscheidung von Klassen und nicht wie bei der unüberwachten Adaption des Erkenners mithilfe des gesamten Erkennermodells und damit anhand der kompletten Merkmalsvektorfolge. Sind die Freiheitsgrade der Anpassung des generischen Modells zu groß, können die Gaußdichten zu stark im Modellraum „verdreht" werden. Schränkt man die Transformation für die Anpassung auf eine Verschiebung der

Mittelwerte ein, müssten zumindest mehrere Verschiebungen geschätzt werden, um die gewichtete Zusammensetzung bei der Akustischen Transformation auszunutzen. Andernfalls werden alle Mittelwerte μ_m der Gaußdichten gleich transformiert ($\tilde{\mu}_m = \mu_m + \mathbf{b}$) und die Abbildung der Muster durch die Akustische Transformation würde ebenfalls dieser konstanten Verschiebung entsprechen.

Mit einem neuen Verfahren soll durch die Akustische Transformation vor allem die Störung durch das additive Rauschen in der Aufnahme der jeweiligen Äußerung kompensiert werden. Im *Log*-Spektral-Bereich, in dem sich Sprache gut akustisch modellieren lässt, ist der Zusammenhang mit dieser Art Störung nicht mehr additiv und kann daher nicht mit einer festen linearen Transformation für alle Merkmalsvektoren ausgeglichen werden. Daher wird bei der Herleitung eines korrespondierenden Modells von den Techniken der Modellkombination (PMC) Gebrauch gemacht. Auf diese Weise werden sowohl die Mittelwerte als auch die Varianzen jeder einzelnen Gaußdichte gesondert angepasst.

Normalerweise wird die PMC (siehe Abschnitt 6.3) zur Anpassung des akustischen Modells des Erkenners benutzt. Dies ist aber bei einer großen Anzahl von Zustandsmodellen für phonetische Klassen sehr rechenaufwändig. Im vorliegenden Fall (siehe Beschreibung des Basiserkenners) müssten 2.500 mal 32, also 80.000 Gaußdichten durch dieses Verfahren neu berechnet werden. Dazu kommen noch einige gravierende Einschränkungen bei der Auswahl der Merkmale, die für einen solchen Erkenner benutzt werden können. Ziel dieser Arbeit ist jedoch eine robuste Erkennung in einer sich ständig ändernden Umgebung. Diese ist nur durch die aktuell zu erkennende Sprachaufnahme repräsentiert und die Anpassung oder Kompensationsmaßnahme soll daher für jede Äußerung individuell erfolgen.

Bei dem hier vorgestellten Verfahren wird die Modellkombination für ein einfaches akustisches Hilfsmodell λ verwendet. Es besitzt nur eine phonetische Klasse (generisches Modell mit M impliziten Unterklassen repräsentiert durch die Gaußdichten) für logarithmierte spektrale Merkmale. Aus der vorliegenden Äußerung wird ein Modell für die additive Störung gewonnen und mit dem Hilfsmodell λ für ungestörte Sprache kombiniert. Mit dem kombinierten Modell $\tilde{\lambda}$ für gestörte Sprache lässt sich dann eine Akustische Transformation (AM) der gestörten Merkmalsvektoren vornehmen, weshalb dieser Ansatz vom Autor **MAM** (engl. ***M**odel Combination Based **A**coustic **M**apping*) genannt wurde.

Nach einer Beschreibung des Umgebungsmodells im nächsten Abschnitt werden die einzelnen Schritte eines darauf basierenden Verfahrens beleuchtet. Dabei werden die Abhängigkeiten von den relevanten Störeinflüssen berücksichtigt, um zu einer generellen Umgebungskompensation zu gelangen. Dies betrifft neben dem Rauschen insbesondere den Kanal, der bei der Schätzung des Geräuschmodells ebenfalls eine Rolle spielt. Das Basisverfahren wird dabei gleich um die Möglichkeit einer **integrierten Sprechernormierung** erweitert.

Grundsätzliches Vorgehen

Bevor jedoch das Verfahren aus mathematischer Sicht beleuchtet wird, soll hier anhand eines konkreten Beispieles die prinzipielle Vorgehensweise des neuen Verfahrens MAM veranschaulicht werden. **Abbildung 9.2** zeigt das Zeitsignal einer durch Rauschen gestörten Sprachaufnahme, die durch den Spracherkenner verarbeitet werden soll. Dazu werden Sprachrahmen entnommen und Merkmale dieser Rahmen berechnet. In der Abbildung ist ein spektraler Merkmalsvektor des verrauschten Sprachsignals zu sehen. Da für dieses Beispiel das Rauschen nur künstlich zugemischt wurde, ist im selben Bild ebenfalls eine ungestörte Version des Spektrums dargestellt. Wie man sieht, sind durch die Störung nur bestimmte Teilbereiche betroffen, die sich eventuell durch Wissen über die Charakteristik des Rauschens und über das „Aussehen" typischer Sprachspektren rekonstruieren ließen.

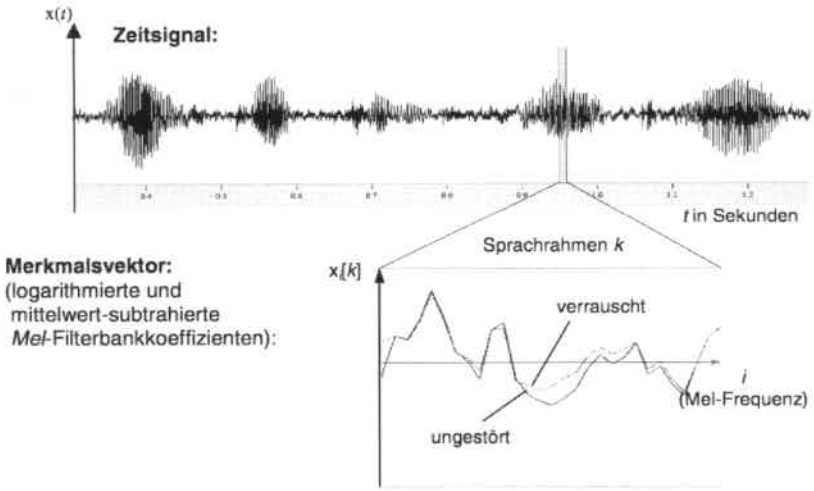


Abbildung 9.2: Ein durch additives Rauschen gestörter Merkmalsvektor.

Die MAM versucht genau dies zu erreichen. Aus dem gestörten Signal wird mittels einer Rauschschätzung und eines Modells für ungestörte Sprache ein Schätzwert für ungestörte Sprache ermittelt. Der Schätzwert wird im *log*-Spektralbereich ermittelt, kann dann aber noch beliebig weiterverarbeitet werden und dient letztendlich als Merkmalsvektor für den Spracherkenner. Das Verfahren läuft grob in zwei Schritten ab:

Schritt 1:

Zunächst werden die durch die Störung bedingten Veränderungen simuliert. Aus der aktuellen Sprachaufnahme werden Geräuschproben entnommen und daraus ein einfaches Geräuschmodell mit einer Gaußdichte im *log*-Spektralbereich erstellt. Dieses wird mit einem zuvor trainierten Hilfsmodell (so genannt, da nicht das akustische Modell des Erkenners verwendet wird) für ungestörte Sprache kombiniert. Es entsteht ein Modell für gestörte

Sprache, dessen Prototypen (Mittelwertvektoren der Gaußdichten) mit denen des ursprünglichen Hilfsmodells korrespondieren.

Schritt 2:

Das neu berechnete Modell beschreibt durch seine Mischverteilung aus M Gaußdichten verrauschte Sprache wie sie in der aktuellen Aufnahme vorliegt. Durch Auswertung der Wahrscheinlichkeitsdichten ergibt sich für den Merkmalsvektor eines jeden Sprachrahmens k eine Zuordnung $p_m[k]$ zu den Prototypen. Hierbei geht unter anderem auch die Varianz des Rauschens ein, da sie bei der Berechnung der Modellparameter (Mittelwert und Varianz der Gaußdichten) eine Rolle spielt. Bei dem hier gezeigten Beispiel ergeben sich oft für jeweils einen einzelnen Prototypen m große Werte für $p_m[k]$, d.h. nahe 1 (**Abbildung 9.3**). Prinzipiell erhält man jedoch eine Verteilung über alle Prototypen. Aus dieser Zuordnung und der Korrespondenz mit den Prototypen für ungestörte Sprache gewinnt man mittels der oben beschriebenen „Akustischen Transformation" einen Schätzwert für den ungestörten Merkmalsvektor. Der dabei benutzte Korrekturvektor ergibt sich aus allen Differenzvektoren der Prototyp-Paare gewichtet mit dem jeweiligen Zuordnungswert $p_m[k]$.

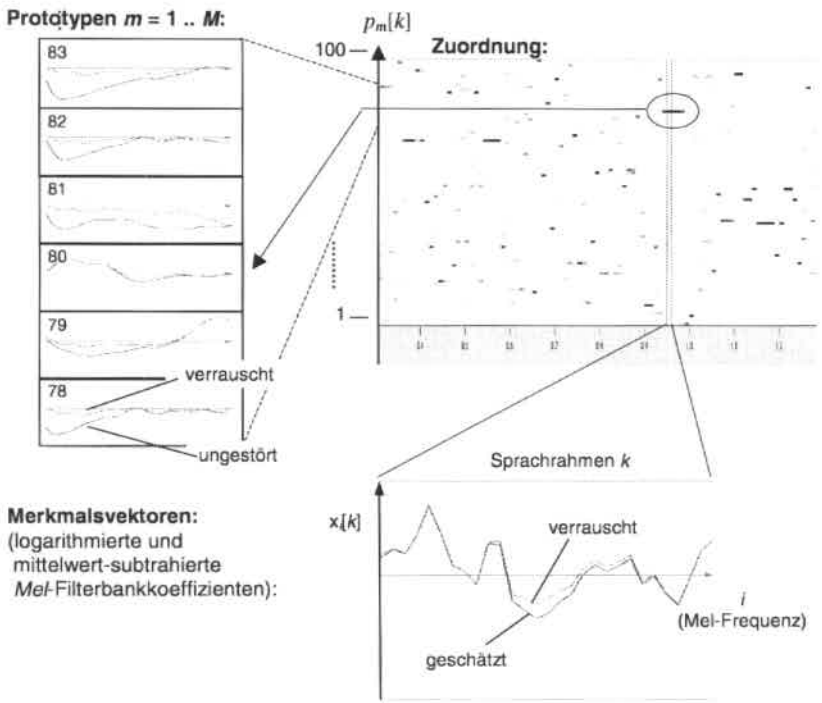


Abbildung 9.3: Zuordnung des verrauschten Merkmalsvektors zu den Prototypen des verrauschten Modells.

9.2 Umgebungskompensation mit MAM

9.2.1 Umgebungsmodell

Das hier vorgestellte Umgebungsmodell stellt eine Erweiterung des klassischen Kanalmodells dar (Abschnitt 8.2, S. 129). Es werden sowohl der Kanal mit Rauschen als auch Sprecher mit unterschiedlichen Vokaltraktlängen (VTL) betrachtet. Alle drei Einflüsse werden hier als Störung (und unterschiedlich zu den Trainingsbedingungen) aufgefasst und sind daher mit „~" gekennzeichnet. Die sich daraus ergebende Beziehung des i-ten Koeffizienten des beobachteten (gestörten) Merkmalsvektors $\tilde{\mathbf{x}}[k]$ zum ungestörten und normierten Sprachanteil $\mathbf{s}[k]$ im Spektralbereich (hochgestelltes „s") ist in **Abbildung 9.4** wiedergegeben.

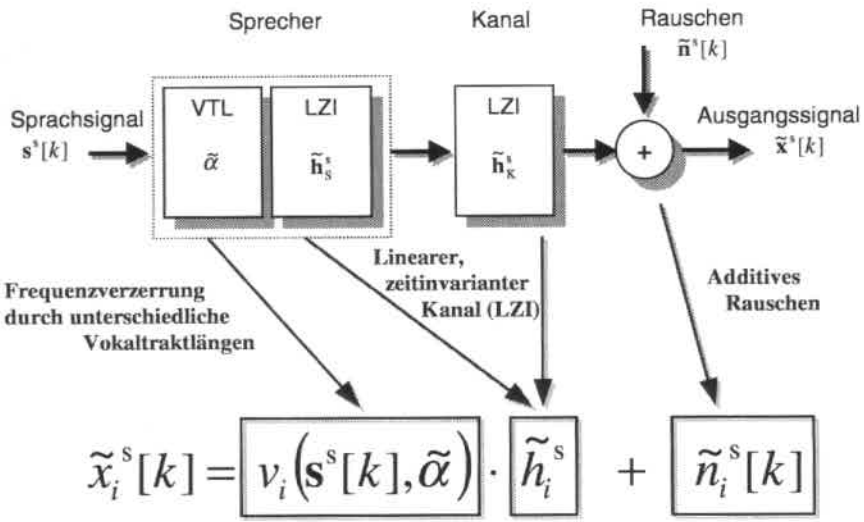


Abbildung 9.4: Umgebungsmodell (erweitertes Kanalmodell).

Die Störungen haben dabei die folgenden Eigenschaften:

- Wechselnder, aber für eine Äußerung konstanter Kanal $\tilde{\mathbf{h}}$. Er setzt sich zusammen aus den linearen, zeitinvarianten Eigenschaften des eigentlichen Übertragungskanals und des Sprecherkanals (siehe S. 131).
- Additives Rauschen $\tilde{\mathbf{n}}[k]$. Besteht aus allen additiven Störkomponenten, die hierin zusammengefasst sind. Es ist für verschiedene Analyserahmen k unterschiedlich, seine Eigenschaften lassen sich aber im *Log*-Spektralbereich mit einer Gaußdichte (Mittelwert und Varianz) beschreiben (vergleiche **Abbildung 8.5**, S. 134).

- Unbekannter Frequenz-Verzerrungsfaktor $\tilde{\alpha}$ (bedingt durch unterschiedliche Vokaltraktlängen). $v_i(\mathbf{s}^s, \tilde{\alpha})$ bezeichnet dabei eine Verzerrungsfunktion in Abhängigkeit des Faktors $\tilde{\alpha}$, die den i-ten Koeffizienten aus dem normierten Spektrum $\mathbf{s}$ erzeugt.

9.2.2 Praktische Umsetzung

Abbildung 9.5 zeigt eine Übersicht der Verfahrensschritte, an denen sich der Leser grob orientieren kann. Die einzelnen Schritte und Überlegungen dazu werden nachfolgend beschrieben. Ein hochgestelltes „l" steht für die logarithmierten Merkmale, ein zusätzliches „α" markiert einen von der Frequenzverzerrung abhängigen Wert.

- **Voraussetzungen:**

  Gegeben seien die K gestörten Merkmalsvektoren einer zu erkennenden Aufnahme. Für deren Koeffizienten gelte das oben beschriebene Umgebungsmodell mit (für die Dauer der Aufnahme) zeitlich konstantem Kanal $\tilde{\mathbf{h}}^s[k] = \tilde{\mathbf{h}}^s$:

  $$\tilde{x}_i^s[k] = \tilde{h}_i^s \cdot v_i\left(\mathbf{s}^s[k], \tilde{\alpha}\right) + \tilde{n}_i^s[k] \quad \text{mit} \quad k = 1 \ldots K \tag{9.11}$$

  Das zur Erkennung verwendete System benutzt ein akustisches Modell (z.B. mit phonetischen Klassen), das auf einer homogenen Sprachdatenbank trainiert wurde. Die Trainingsaufnahmen sind kaum verrauscht, Kanal und Sprecher wurden während des Trainings normiert. Daneben wird für Schritt 3 dieses Verfahrens ein generisches Modell $\lambda_{\mathbf{s}}^{\,l}$ für die Sprachmuster $\mathbf{s}$ der Trainingsmenge im *Log*-Spektralbereich gebraucht.

- **Schritt 1:** Sprechernormierung durch VTLN-Frequenzverzerrung

  Basierend auf einer Modellannahme für den Vokaltrakt soll dessen Einfluss durch die Funktion $\hat{v}_i^{-1}\left(\tilde{\mathbf{x}}^s[k], \hat{\alpha}\right)$ kompensiert werden. Der Schätzwert $\hat{\alpha}$ für den Verzerrungsfaktor α ist zunächst noch unbekannt. Er wird erst in Schritt 4 durch eine Gittersuche mit dem Maximum-Likelihood-Kriterium bestimmt. Daher nimmt α zunächst N_α verschiedene Werte an.

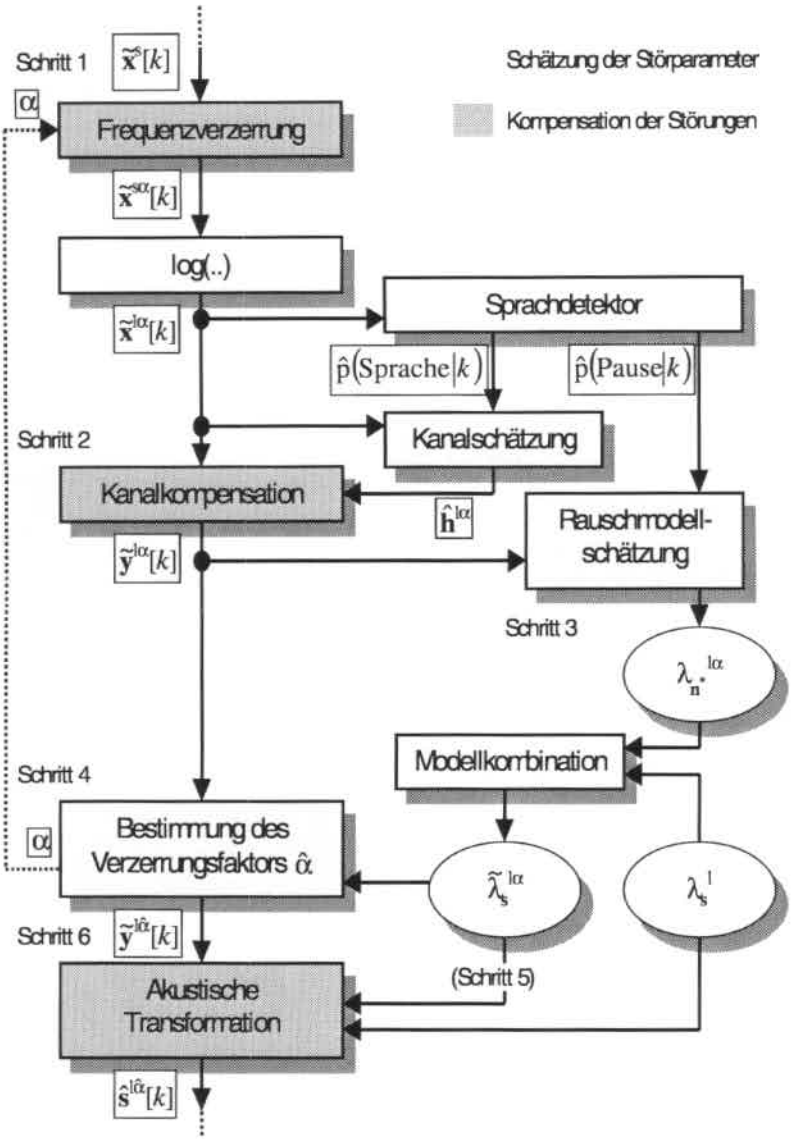


Abbildung 9.5: Überblick der sprechernormierenden Modellkombinationsbasierten Akustischen Transformation (MAM).

Wie in Kapitel 7 Sprechernormierung beschrieben, wird das Spektrum jeweils entlang der Frequenzachse näherungsweise linear verzerrt. Die Koeffizienten des verzerrten Spektrums $\tilde{x}_i^{s\alpha}[k]$ werden durch lineare Interpolation des unverzerrten Spektrums bestimmt, was man als Multiplikation mit einer Matrix $\mathbf{V}^{\alpha}$ schreiben kann. Ein verzerrter Koeffizient ergibt sich jeweils aus der Multiplikation mit der i-ten Zeile $\mathbf{V}_{(i)}{}^{\alpha}$. Die Werte dieser Zeile sind überwiegend Null, bis auf die Interpolationsgewichte für benachbarte Koeffizienten von $\tilde{\mathbf{x}}^{s}[k]$ (vergleiche **Abbildung 7.4**, S. 111).

$$\tilde{x}_i^{s\alpha}[k] = \hat{v}_i^{-1}\left(\tilde{\mathbf{x}}^{s}[k],\alpha\right) = \mathbf{V}_{(i)}{}^{\alpha} \cdot \tilde{\mathbf{x}}^{s}[k] \tag{9.12}$$

$$\begin{aligned}
\tilde{\mathbf{x}}^{s\alpha}[k] &= \mathbf{V}^{\alpha} \cdot \tilde{\mathbf{x}}^{s}[k] \\
&= \mathbf{V}^{\alpha} \cdot \left(\begin{pmatrix} \vdots \\ \tilde{h}_i^{s} v_i\left(\mathbf{s}^{s}[k],\tilde{\alpha}\right) \\ \vdots \end{pmatrix} + \tilde{\mathbf{n}}^{s}[k] \right) \\
&= \mathbf{V}^{\alpha} \cdot \begin{pmatrix} \ddots & 0 & \\ & \tilde{h}_i^{s} & \\ 0 & & \ddots \end{pmatrix} \cdot \begin{pmatrix} \vdots \\ v_j\left(\mathbf{s}^{s}[k],\tilde{\alpha}\right) \\ \vdots \end{pmatrix} + \mathbf{V}^{\alpha} \cdot \tilde{\mathbf{n}}^{s}[k] \\
&= \mathbf{V}^{\alpha} \cdot \begin{pmatrix} \ddots & 0 & \\ & \tilde{h}_i^{s} & \\ 0 & & \ddots \end{pmatrix} \cdot \left(\mathbf{V}^{\alpha}\right)^{-1} \cdot \mathbf{V}^{\alpha} \cdot \begin{pmatrix} \vdots \\ v_j\left(\mathbf{s}^{s}[k],\tilde{\alpha}\right) \\ \vdots \end{pmatrix} + \tilde{\mathbf{n}}^{s\alpha}[k] \\
&= \mathbf{H}^{\alpha} \cdot \mathbf{V}^{\alpha} \cdot \begin{pmatrix} \vdots \\ v_j\left(\mathbf{s}^{s}[k],\tilde{\alpha}\right) \\ \vdots \end{pmatrix} + \tilde{\mathbf{n}}^{s\alpha}[k]
\end{aligned} \tag{9.13}$$

Die Einflüsse von Kanal und Rauschen konnten aufgrund der linearen Operation wieder getrennt werden. Der Rauschanteil wurde durch das verzerrte Rauschspektrum mit den Koeffizienten $\tilde{n}_i^{s\alpha}[k]$ ersetzt, die auf diese Weise wieder direkt mit den gestörten Merkmalskoeffizienten $\tilde{x}_i^{s\alpha}[k]$ in Beziehung gesetzt werden können. Für den Kanal gelingt dies nicht. Aufgrund der Eigenschaften von $\mathbf{V}^{\alpha}$ kann man jedoch in der Praxis $\mathbf{H}^{\alpha}$ durch eine Diagonalmatrix annähern und gelangt zur folgenden Beziehung der Koeffizienten nach der Frequenzverzerrung:

$$\tilde{x}_i^{s\alpha}[k] \approx \tilde{h}_i^{s\alpha} \cdot \hat{v}_i^{-1}\left(\begin{pmatrix} \vdots \\ v_j\left(\mathbf{s}^{s}[k],\tilde{\alpha}\right) \\ \vdots \end{pmatrix}, \alpha \right) + \tilde{n}_i^{s\alpha}[k] \tag{9.14}$$

Mit einem geeigneten Schätzwert $\alpha = \hat{\alpha}$ für die Frequenzverzerrung können also die Schwankungen bedingt durch unterschiedliche Vokaltraktlängen ausgeglichen werden. Die noch folgende Kanal- und Rauschkompensation muss abhängig von α erfolgen.

- **Schritt 2:** Kanalkompensation für $\tilde{\mathbf{h}}$ mit sprachbasierter CMS

  Die Kanalkompensation erfolgt nach dem Zusammenfassen der Spektralkoeffizienten

durch eine Filterbank und Übergang in den logarithmischen Spektralbereich (hier als $\tilde{x}^{l\alpha}[k]$ bezeichnet). Sie wird **vor** der Geräuschkompensation durch MAM durchgeführt, um den Modellierungsaufwand für das in Schritt 3 verwendete Hilfsmodell zu verringern. Dieses soll ein bezüglich der Umgebung normiertes Modell der Sprachmuster darstellen, um so erstens mit wenig Parametern auszukommen und zweitens als Referenz für ungestörte Sprache zu dienen. Da die Kanalschätzung nicht von den Sprachpausen abhängen soll, wird eine sprachbasierte CMS verwendet (siehe Abschnitt 8.4.4). Ein einfaches Energiemaß als Näherung für p(Sprache|k) ist hierfür ausreichend, da die Muster mit einem entsprechend hohen Energiewert am wenigsten durch das additive Rauschen gestört sind.

$$\hat{\mathbf{h}}^{l\alpha} = \frac{\sum_{k=1}^{K} \hat{p}(\text{Sprache} \mid k) \cdot \tilde{\mathbf{x}}^{l\alpha}[k]}{\sum_{k=1}^{K} \hat{p}(\text{Sprache} \mid k)} \tag{9.15}$$

$$\tilde{y}_i^{l\alpha}[k] = \tilde{x}_i^{l\alpha}[k] - \hat{h}_i^{l\alpha} = \log\left(\frac{\tilde{h}_i^{s\alpha}}{\hat{h}_i^{s\alpha}} \cdot \hat{v}_i^{-1}\left(\begin{pmatrix} \vdots \\ v_j\left(\mathbf{s}^s[k], \tilde{\alpha}\right) \\ \vdots \end{pmatrix}, \alpha\right) + \frac{\tilde{n}_i^{s\alpha}[k]}{\hat{h}_i^{s\alpha}}\right) \tag{9.16}$$

$$\text{mit} \quad \hat{h}_i^{s\alpha} = \exp\left(\hat{h}_i^{l\alpha}\right)$$

- **Schritt 3:** Berechnung eines korrespondierenden Modells für die gestörte Sprache

  Es wird zuvor einmalig eine einfache Mischverteilung $\lambda_s^{\,1}$ für ungestörte Sprache auf den Trainingsdaten berechnet (siehe Voraussetzungen oben). Als Merkmalsvektoren werden dazu $\mathbf{y}^{l\hat{\alpha}}[l]$ verwendet (wobei $\hat{\alpha}$ der Verzerrungsfaktor des jeweiligen Trainingssprechers ist und l alle Trainingsbeispiele indiziert). Das Rauschmodell wird hingegen auf der aktuellen Aufnahme bestimmt. Für jede Frequenzverzerrung wird dabei eine eigene kanalkompensierte Version erzeugt:

  3.1: Modell für das kanalkompensierte, frequenzverzerrte Rauschen $n_i^*[k] = \frac{\tilde{n}_i^{s\alpha}[k]}{\hat{h}_i^{s\alpha}}$:

  $$\lambda_{n^*}^{l\alpha} = \left\{\boldsymbol{\mu}_{n^*}^{l\alpha}, \mathbf{C}_{n^*}^{l\alpha}\right\} \quad \text{mit}$$

  $$\boldsymbol{\mu}_{n^*}^{l\alpha} = \frac{\sum_{k=1}^{K} \hat{p}(\text{Pause} \mid k) \cdot \tilde{\mathbf{y}}^{l\alpha}[k]}{\sum_{k=1}^{K} \hat{p}(\text{Pause} \mid k)}, \tag{9.17}$$

  $$\mathbf{C}_{n^*}^{l\alpha} = \frac{\sum_{k=1}^{K} \hat{p}(\text{Pause} \mid k) \cdot \left(\tilde{\mathbf{y}}^{l\alpha}[k] - \boldsymbol{\mu}_{n^*}^{l\alpha}\right) \cdot \left(\tilde{\mathbf{y}}^{l\alpha}[k] - \boldsymbol{\mu}_{n^*}^{l\alpha}\right)^T}{\sum_{k=1}^{K} \hat{p}(\text{Pause} \mid k)} \tag{9.18}$$

3.2: Kombination des Rauschmodells mit dem Hilfsmodell für ungestörte Sprache $\lambda_s^{\,l}$:

Dazu werden die Modelle $\lambda_s^{\,l}$ und $\lambda_n^{\,l\alpha}$ zunächst in den linearen Spektralbereich transformiert und dort kombiniert. Das kombinierte Modell für verrauschte Sprache wird schließlich wieder in den logarithmischen Bereich überführt (siehe **Tabelle 6.1**, S. 93):

$$\tilde{\lambda}^{l\alpha} = PMC\left(\lambda_s^{\,l}, \lambda_n^{\,l\alpha}\right) \tag{9.19}$$

- **Schritt 4:** Bestimmung des Verzerrungsfaktors $\hat{\alpha}$

Es liegt nun für jeden der betrachteten Verzerrungsfaktoren ein Modell $\tilde{\lambda}^{l\alpha}$ für verrauschte Sprache vor. Man beachte, dass hierfür nicht der im Modell enthaltene Sprachanteil, sondern lediglich der kanalnormierte Rauschanteil frequenzverzerrt wurde, damit er zu den vorliegenden Merkmalsvektoren passt. Die entsprechende Version $\tilde{\mathbf{Y}}^{l\alpha} = \tilde{\mathbf{y}}^{l\alpha}[1] \ldots \tilde{\mathbf{y}}^{l\alpha}[K]$ der verzerrten Merkmalsvektorfolgen kann nun mit diesem Modell für sprechernormierte Sprache bewertet werden. Ein für den aktuellen Sprecher geeigneter Verzerrungsfaktor $\hat{\alpha}$ wird wie bei der ML-VTLN (Abschnitt 7.4, S. 116) über die Auswertung der Likelihood bestimmt:

$$\hat{\alpha} = \arg\max_{\alpha} \; p\left(\tilde{\mathbf{Y}}^{l\alpha} \mid \tilde{\lambda}^{l\alpha}\right) \tag{9.20}$$

Alternativ könnte man auch die Merkmalsvektoren und damit das Rauschmodell unverzerrt lassen und im Gegenzug das normierte Sprachmodell $\lambda_s^{\,l}$ dem aktuellen Sprecher anpassen. Dazu könnten die Mittelwerte des Modells nach der Transformation in den Spektralbereich mit einer Funktion $\hat{v}\left(\mu^s{}_m, \alpha\right)$ verzerrt werden. Allerdings handelt es sich bei den Mittelwerten $\mu^s{}_m$ um *Mel*-Spektralkoeffizienten, bei denen die feiner aufgelösten Spektralkoeffizienten, die hier sonst zur Verzerrung benutzt werden, zusammengefasst wurden. Eine Frequenzverzerrung könnte also nur sehr ungenau vorgenommen werden. Mittels des so entstandenen kombinierten Modells $\tilde{\lambda}_2^{\,l\alpha}$ für Sprecher mit dem Verzerrungsfaktor α könnte die Folge $\tilde{\mathbf{Y}}^{l}$ bewertet werden. Als Schätzwert ergibt sich dann:

$$\hat{\alpha}_2 = \arg\max_{\alpha} \; p\left(\tilde{\mathbf{Y}}^{l} \mid \tilde{\lambda}_2^{\,l\alpha}\right) \tag{9.21}$$

Für die Experimente wurde nur die erste Methode eingesetzt.

Im Gegensatz zu den Experimenten mit ML-VTLN in Kapitel 7, bei denen das akustische Modell des Erkenners verwendet wurde, wird hier das Hilfsmodell der aktuellen Geräuschumgebung angepasst. Die Verwendung eines generischen Modells (für alle phonetische Klassen) hat sowohl Vorteile als auch Nachteile. Ein Vorteil ist, dass keine phonetische Klassenzuordnung der Merkmalsvektoren und damit kein erster Erkennungslauf gebraucht wird. Diese Zuordnung lässt jedoch andererseits eine bessere Schätzung des

Verzerrungsfaktors mit wenigen Mustern zu. Vorraussetzung ist allerdings, dass diese Zuordnung auch zuverlässig ist. Wie oben gezeigt wurde, ist dies bei stark verrauschten Daten nicht der Fall.

- **Schritt 5 (optional):** Adaption des Modells $\tilde{\lambda}^{1\hat{a}}$, z.B. mit MLLR.

  Das Hilfsmodell wurde bereits über die Modellkombination an die aktuelle Rauschsituation angepasst und könnte nun durch Vergleich mit den vorliegenden Merkmalsvektoren der Äußerung noch weiter angepasst werden. In Abschnitt 9.1.3 wurden jedoch schon Gründe genannt, warum dies für ein generisches Modell problematisch sein kann. In Vorversuchen vielen die Ergebnisse mit einer weiteren Anpassung des Hilfsmodells mit MLLR sehr viel schlechter aus als ohne, weshalb dieser Schritt bei den unten vorgestellten Experimenten nicht durchgeführt wurde.

- **Schritt 6:** Akustische Transformation

  Durch den letzten Schritt wird mithilfe der Akustischen Transformation aus dem Merkmalsvektor $\tilde{\mathbf{y}}^{1\hat{a}}[k]$ (sprecher- und kanalnormiert, aber mit frequenzverzerrtem Rauschen) mit dem Hilfsmodell $\tilde{\lambda}^{1\hat{a}}$ (normierte Sprache, frequenzverzerrtes Rauschen) ein dem korrespondierenden Modell λ_s^{1} (normierte, ungestörte Sprache) entsprechender Merkmalsvektor $\hat{\mathbf{s}}^{1}[k]$ geschätzt.

$$\hat{\mathbf{s}}^{1}[k] = AM\left(\tilde{\mathbf{y}}^{1\hat{a}}[k]; \lambda_s^{1}, \tilde{\lambda}^{1\hat{a}}\right) \tag{9.22}$$

9.3 Versuchsergebnisse

9.3.1 MAM ohne Sprechernormierung

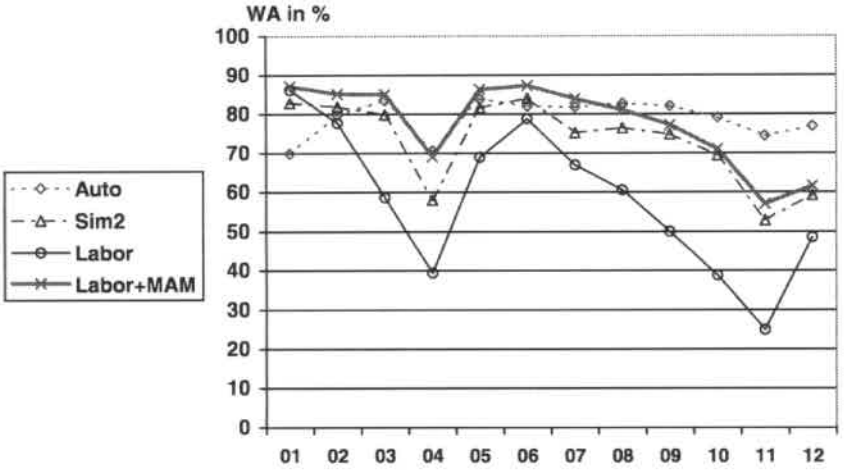
Parameterwahl und Training des Hilfsmodells

Im Gegensatz zu vielen anderen Methoden hat das gesamte oben beschriebene Verfahren kaum offene Parameter, die zum Ausgleich von Modellierungsschwächen experimentell für besondere Einsatzbedingungen bestimmt werden müssten. Wie bei anderen Kompensationsverfahren betrifft dies den Sprach-Pause-Detektor, der hier zur Schätzung des Rauschmodells und des Kanals gebraucht wird. Die Parametereinstellungen sind aber relativ unkritisch, da in beiden Fällen eine exakte Klassifikation aller Muster nicht notwendig ist. Für die Rauschschätzung sollte lediglich sichergestellt sein, dass dafür keine Sprachmuster verwendet werden. Bei der Auswahl kann man daher auf die sicher erkannten Pausenregionen zurückgreifen. Zur Schätzung des Kanals reicht ein einfaches Energiemaß, um damit relativ sicher einen Teil der Sprachmuster zu erhalten.

Davon abgesehen muss nur noch festgelegt werden wie viele Gaußdichten das akustische Hilfsmodell haben soll. Verschieden große generische Modelle mit diagonalen Kovarianzmatrizen wurden deshalb mithilfe der gleichen Trainingsdaten erstellt, die auch für den Laborerkenner verwendet wurden. Da die Zahl der Modellparameter verglichen mit dem akustischen Modell des Erkenners in allen betrachteten Fällen sehr viel kleiner gewählt wurde, reichte ein wesentlich kleinerer Teil dieser Daten aus. Als Merkmale wurden die logarithmierten *Mel*-Spektralkoeffizienten mit sprachbasierter Kanalkompensation verwendet. Der Laborerkenner verwendet darauf aufbauende Merkmale, die noch die cepstrale Analyse, dynamische Koeffizienten und eine LDA-Transformation mit einschließt (vergleiche Abschnitt 4.2, S. 60). Für die Experimente mit der sprechernormierenden MAM (s.u.) wurden die Trainingsmuster mit den Verzerrungsfaktoren der LD-VTLN sprechernormiert. Diese Faktoren stimmen mit denen überein, die auch beim Training des akustischen Modells des VTLN-Spracherkenners benutzt wurden.

Beim Vergleich zwischen 10 und 100 Gaußdichten für das Hilfsmodell ergaben sich noch merkliche Unterschiede bei der Erkennungsleistung. Die relative Fehlerreduktion lag hier um 10 %, bei starker Störung sogar über 20 %. Eine weitere Erhöhung auf 1.000 Gaußdichten brachte keine weitere Verbesserungen mehr. Die folgenden Ergebnisse beruhen daher alle auf einem akustischen Hilfsmodell mit 100 Gaußdichten. Mit der 2DCMS, die zwei explizite Klassen verwendet, wurden (wie in Abschnitt 8.5.2, S. 154 gezeigt) bereits deutliche Verbesserungen der Erkennungsleistung erzielt. Die Wortfehlerrate des Basissystems konnte dabei bis zu 40 % reduziert werden. Durch das neue Verfahren wurden diese Ergebnisse nochmals deutlich übertroffen.

Vergleich mit den Referenzsystemen



Kategorie	01	02	03	04	05	06	07	08	09	10	11	12
System	aus	Blinker	Fenster	Lstark	Lmittel	0km/h	25km/h	50km/h	75km/h	100km/h	125km/h	beschl.
Basiserkenner (WA in %)	86,1	77,7	58,6	39,4	68,9	78,7	66,9	60,5	50,0	38,7	25,0	48,5
mit MAM (WA in %)	87,1	85,1	85,0	69,1	86,3	87,3	83,9	81,0	77,2	71,0	57,0	61,5
ER relativ	7%	33%	64%	49%	56%	40%	51%	52%	54%	53%	43%	25%

Abbildung 9.6: Ergebnisse der MAM ohne Sprechernormierung

Abbildung 9.6 zeigt die Erkennungsraten des Basissystems mit MAM (zunächst ohne Sprechernormierung) im Vergleich mit den Referenzsystemen. Die Fehlerreduktionen gegenüber dem Basiserkenner (Labor) erreichen bis über 60 % und fallen sogar für die ungestörte Kategorie 01 auf etwa 7 % aus. Dieser Erkenner (Labor+MAM), für den überhaupt keine verrauschten Daten im Vorfeld verwendet wurden und der auch (bis auf das Hilfsmodell) nicht neu trainiert werden musste, liefert Ergebnisse, die durchweg besser sind als die des Erkenners Sim2, der mit künstlich verrauschten Daten neu trainiert wurde. Für viele Kategorien werden sogar die Ergebnisse des aufwändig trainierten Autospracherkenners (Auto) erreicht oder übertroffen. Lediglich für starke Fahrgeräusche (Kategorie 08-12) ist letzterer überlegen.

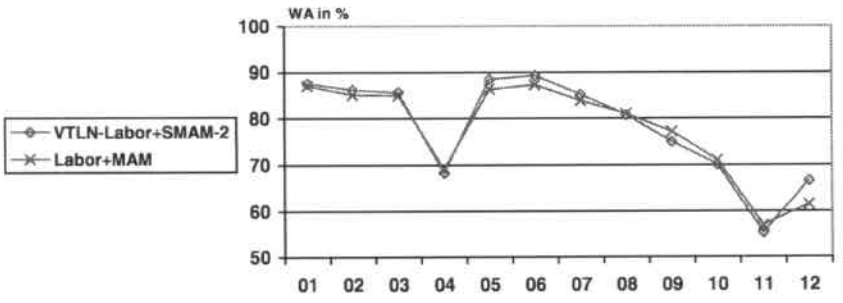
9.3.2 MAM mit Sprechernormierung

Bei Verwendung der integrierten Sprechernormierung erwies sich die Schätzung des Verzerrungsfaktors auf nur einer Äußerung mithilfe des angepassten generischen Modells als nicht sehr zuverlässig. Die Varianz der für einen Sprecher gefundenen Verzerrungsfaktoren ist gegenüber der ML-VTLN mithilfe des Erkennermodells (zumindest für die ungestörten Kategorien) erhöht. Auch mit dem Erkennermodell war die Bestimmung auf Grundlage einer Äußerung schon schwierig und konnte nur über die Auswahl der stimmhaften Merkmalsvektoren zu deutlichen Verbesserungen führen. Dies wäre beim generischen Modell ebenfalls

eine Option, wenn es gelingt die stimmhaften Bereiche des gestörten Signals einzugrenzen. Da der Sprecher sich während einer Anwendung nicht in dem Maß ändert wie die Geräuschkulisse, wurde hier (in Abweichung der sonst gestellten Anforderungen) jedoch ein anderer Weg untersucht.

Von jedem Sprecher stammen in jeder Testkategorie 30 Äußerungen. Bestimmt man den Verzerrungsfaktor nicht für jede Äußerung getrennt, sondern bezieht die vorigen Äußerungen dafür mit ein, pendelt sich der Faktor nach etwa drei bis fünf Sätzen auf einen Wert ein. Einen Vergleich dieses Testverfahrens mit der MAM ohne Sprechernormierung zeigt **Abbildung 9.7** (WA-Bereich vergrößert dargestellt, zwischen 50 und 100 %). Es ergeben sich hierbei Verbesserungen für die etwas weniger gestörten Kategorien, aber auch Verschlechterungen für die stärker gestörten Kategorien.

Das Verhalten ist also ähnlich wie bei der Sprechernormierung in Kapitel 7, bei der die Verzerrungsfaktoren mithilfe des akustischen Modells des Erkenners gefunden wurden. Die Erkennungsgeschwindigkeit ist diesmal jedoch wesentlich schneller, da keine erste Hypothese für die Zuordnung der Muster zu phonetischen Teilmodellen benötigt wird. Da diese erste Hypothese des Erkenners auf einem mittleren und damit meist falschen Verzerrungsfaktor berechnet werden muss, ist eine Kombination folgender Art denkbar: Zunächst erfolgt eine Schätzung des Verzerrungsfaktors durch das Hilfsmodell, mit der dann ein erster Erkennungslauf auf bereits durch die MAM geräuschreduzierten Mustern erfolgen kann. Mithilfe der Hypothese wird über das Erkennermodell ein verbesserter Verzerrungsfaktor bestimmt, der bei der MAM für eine zweite Erkennung berücksichtigt wird. Der Nachteil des erhöhten Rechenzeitbedarfs bleibt jedoch auf diese Weise erhalten, weshalb diese Möglichkeit nicht untersucht wurde.



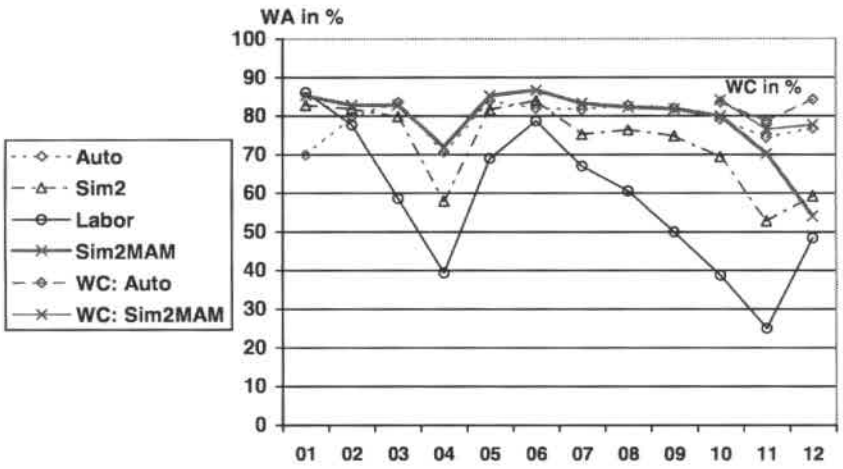
Kategorie	01	02	03	04	05	06	07	08	09	10	11	12
System	aus	Blinker	Fenster	Lstark	Lmittel	0km/h	25km/h	50km/h	75km/h	100km/h	125km/h	beschl.
Labor+MAM (WA in %)	87,1	85,1	85,0	69,1	86,3	87,3	83,9	81,0	77,2	71,0	57,0	61,5
VTLN-Labor+SMAM-2 (WA	87,6	86,2	85,6	68,3	88,5	89,3	85,2	80,8	75,0	70,1	55,4	66,6
ER relativ	4%	7%	4%	-3%	16%	16%	8%	-1%	-10%	-3%	-4%	13%

Abbildung 9.7: Ergebnisse der MAM mit Sprechernormierung (Training LD, Test ML). Hier wurden die Verzerrungsfaktoren fortlaufend über die Äußerungen eines Sprechers bestimmt.

9.3.3 MAM im Training

Die MAM wurde zwar für den Einsatz in der Erkennungsphase entwickelt, um die Muster dort den Trainingsbeispielen ähnlicher zu machen, sie kann aber auch bereits während der Trainingsphase verwendet werden. Bedingung ist, dass für die Erstellung des Hilfsmodells unverrauschte Daten verwendet werden, die Trainingsdaten für den Erkenner hingegen können verrauscht sein. Die Verbesserungen des Laborerkenners durch MAM ausschließlich während der Erkennung waren bereits beachtlich, obwohl dieser im Training nie Beispiele verrauschter und dann kompensierter Muster vorliegen hatte. Eine interessante Frage soll daher mit einem weiteren Versuch beantwortet werden: Kann man weitere Verbesserungen der Erkennungsleistung erzielen, mit einem System, das bereits mit MAM trainiert wurde und damit im akustischen Modell eventuelle Schätzfehler berücksichtigt?

Zunächst wurde der Laborerkenner lediglich unter Hinzunahme der MAM auf den ungestörten Labordaten neu trainiert. Die Ergebnisse waren nur für wenige Kategorien und dort auch nur unwesentlich verbessert. Verwendet man jedoch zusätzlich Rauschaufnahmen, die man wie für den Erkenner Sim2 beim Training zufällig gewichtet hinzu addiert, ergeben sich größere Gewinne (**Abbildung 9.8**). Im Vergleich zum Basissystem werden durch die zusätzliche Varianz in den Trainingsdaten (wie bei Sim2, Kategorie 01) die Ergebnisse für die weniger gestörten Kategorien etwas schlechter. Bei starkem Fahrgeräusch hingegen, erhält man bis zu der hier betrachteten Maximalgeschwindigkeit 125 km/h viel bessere Ergebnisse.



Kategorie / System	01 aus	02 Blinker	03 Fenster	04 Lstark	05 Lmittel	06 0km/h	07 25km/h	08 50km/h	09 75km/h	10 100km/h	11 125km/h	12 beschl.
Basiserkenner (WA in %)	86,1	77,7	58,6	39,4	68,9	78,7	66,9	60,5	50,0	38,7	25,0	48,5
mit MAM (WA in %)	85,2	82,8	82,8	71,9	85,3	86,7	83,2	82,2	81,8	79,9	70,2	54,1
ER relativ	-6%	23%	58%	54%	53%	38%	49%	55%	64%	67%	60%	11%

Abbildung 9.8: Ergebnisse der MAM für Training auf simulierten Daten ohne Sprechernormierung.

Nun sind die erreichten Wortakkuratheiten fast überall besser oder vergleichbar mit denen des Autospracherkenners. Eine Ausnahme stellt die Kategorie 12 dar (beschleunigen), die deutlich abweicht. Betrachtet man allerdings die Wortkorrektraten WC (die oben rechts in **Abbildung 9.8** mit aufgenommen wurden) erkennt man, dass diese weit weniger gravierend einbrechen, sich sogar gegenüber Kategorie 11 verbessern. Die Fehler beruhen also vor allem auf Einfügungen. Das verwendete Geräuschmodell beruht auf einer Schätzung des Mittelwertes und der Varianz auf wenigen Pausenrahmen und kann wohl das sich ändernde Geräusch nicht gut genug erfassen. So verbleibt am Ende der Aufnahmen einiges Restrauschen und dadurch resultieren Einfügefehler. Die im Training addierten Geräuschaufnahmen sind bei relativ konstanter Geschwindigkeit aufgenommen, der Fall starker Beschleunigung wurde daher nicht trainiert. In der Praxis ist dieser Fall auch weniger relevant, da die wenigsten Fahrer bei einer starken Beschleunigung gleichzeitig die Spracherkennung benutzten werden.

Die sonstigen Kategorien liegen fast alle über 80 % Wortakkuratheit. Lediglich bei sehr lauten Störgeräuschen, wie starke Lüftung (04) oder bei Tempo 125 km/h (11), verringert sich die Leistung auf etwa 70 %. Erfreulich ist, dass gleichzeitig auch das Ergebnis für die ungestörte (01) und die leicht gestörten Kategorien (05-07) auf dem Niveau des besten Resultates des Basiserkenners liegt, der lange für diese leise Umgebung optimiert wurde.

9.4 Vergleich und Bewertung

Vergleich mit anderen Verfahren für additives Rauschen

Zunächst soll anhand eines Beispiels veranschaulicht werden, wie durch MAM ungestörte Merkmalskoeffizienten rekonstruiert werden und weshalb dies mit Geräuschreduktionsfiltern (Abschnitt 8.4, S. 137), die nur einzelne Spektralbänder betrachten, weniger gut gelingt. Außerdem wird daran gezeigt, welche Gemeinsamkeiten mit der *Missing Features Theory* (Abschnitt 5.2, S. 76) bestehen.

Ein typisches Sprachspektrum zeichnet sich dadurch aus, dass es bestimmte Resonanzstellen (Formanten) aufweist oder sich bestimmte Frequenzbereiche durch eine hohe Leistung hervorheben. Das heißt, während einige energieschwache Koeffizienten durch additives Rauschen sehr stark beeinflusst werden können, bleiben andere relativ ungestört. **Abbildung 9.9** zeigt exemplarisch ungestörte Muster $\mathbf{x}[k]$ mit zwei *Log*-Spektralmerkmalen (z.B. zwei logarithmierte *Mel*-Bänder) von drei Klassen A, B und C. Durch eine im spektralen Bereich additive Störung wurden diese Muster beeinflusst (**Abbildung 9.10**). Wie in **Abbildung 8.8** auf Seite 136 wurden die Muster mit kleinen Werten für einen Koeffizienten $x_i[k]$ nun hin zum entsprechenden Mittelwert des Rauschens $\overline{n}_i$ verschoben und verteilen sich unregelmäßig um ihn.

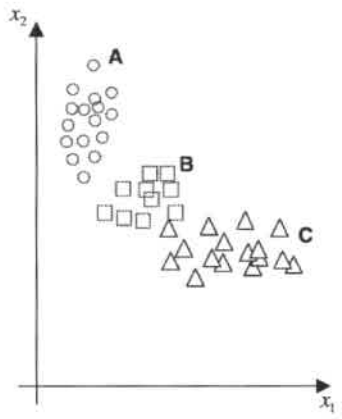


Abbildung 9.9: Ungestörte Muster (2 *Log*-Spektralbänder als Merkmal).

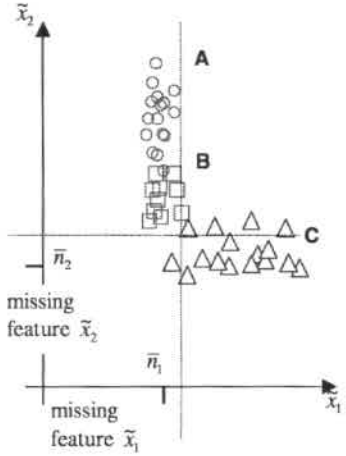


Abbildung 9.10: Durch additives Rauschen gestörte Muster.

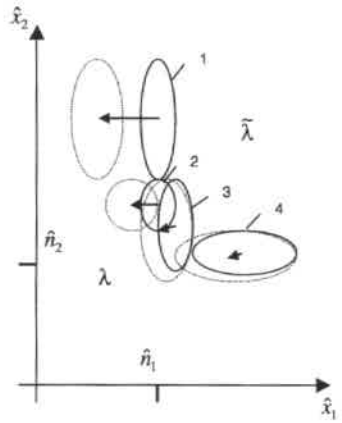


Abbildung 9.11: Geräuschreduktion mittels MAM.

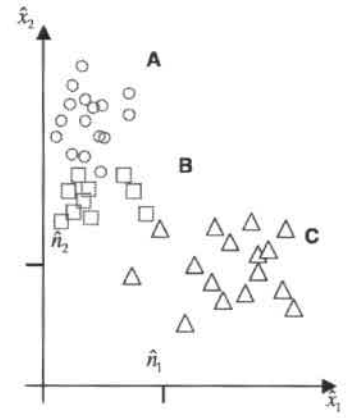


Abbildung 9.12: Geräuschreduktion mittels Spektraler Subtraktion.

In der *Missing Feature Theory* wird ein Schwellwert in Abhängigkeit von einem Schätzwert $\hat{n}_i$ für das mittlere Rauschen bestimmt und alle Koeffizient $\tilde{x}_i[k]$, die unter diesen Schwellwert fallen, als unzuverlässig erklärt. Eine der Möglichkeiten weiter zu verfahren ist, diese fehlenden Koeffizienten anhand der zuverlässigen Koeffizienten dieses Musters zu rekonstruieren (siehe Abschnitt 5.2, S. 76). Dabei spielt die Wahl des Schwellwertes eine

wichtige Rolle. Ist er zu klein gewählt, bleiben viele Koeffizienten gestört und werden darüber hinaus auch zur Rekonstruktion der fehlenden Werte verwendet. Bei einem zu großen Schwellwert bleiben nur sehr wenige zuverlässige Koeffizienten übrig, aus denen alle übrigen geschätzt werden müssen.

Bei der Geräuschreduktion mittels MAM (**Abbildung 9.11**) ist diese explizite Entscheidung nicht erforderlich. Sie erfolgt implizit über eine Zuordnung der gestörten Muster zu den Gaußdichten des Modells $\tilde{\lambda}$ für gestörte Sprachmuster. Im Falle von Gaußdichten mit diagonalen Kovarianzmatrizen (wie sie auch hier verwendet wurden) kann man diese Zuordnung koeffizientenweise aufspalten (vergleiche Gleichung (5.1), S. 75). Für viele gestörte Muster der Klassen A und B ist im Beispiel der aus $\tilde{x}_1$ errechnete Beitrag der Gaußdichten 1 und 2 fast gleich. Die Zuordnung zu einer der beiden Dichten erfolgt über den Beitrag, der sich aus dem – wenn man es so betrachten will – *zuverlässigen* Koeffizient $\tilde{x}_2$ ergibt. Die Rekonstruktion erfolgt hingegen nicht nur anhand der zuverlässigen, sondern unter Zuhilfenahme aller Koeffizienten (d.h. des kompletten Mermalsvektors) mittels des Korrekturvektors, der sich aus den Zuordnungen ergibt.

Bei der Spektralen Subtraktion (**Abbildung 9.12**) erfolgt die Rekonstruktion ohne ein akustisches Modell für ungestörte Sprache und für jeden Koeffizienten getrennt. Anhand einer Rauschschätzung wird daher nichts anderes getan, als die gestörten Koeffizienten in einen bestimmten Bereich abzubilden, der nur durch das vorgegebene Restrauschen begrenzt wird.

Erst durch die Verwendung eines lokal begrenzten SNR-Wertes, wie bei dem Verfahren von Ephraim und Malah, wird auch Information anderer (zeitlich vorangegangener) Koeffizienten in ein Geräuschreduktionsfilter eingebracht. Damit werden Koeffizienten in Pausenregionen, die typischerweise besonders gestört sind, stärker gedämpft als die von Sprachmustern. Ähnlich arbeitet auch die 2DCMS zur Kompensation von Kanal und Rauschen. Die Gewichtung der beiden zuvor ermittelten Korrekturwerte erfolgt abhängig vom gesamten Merkmalsvektor, eventuell sogar (durch Glättung) abhängig von benachbarten Mustern. Vergleicht man diese beiden Verfahren, so liegt nur im Falle der 2DCMS ein wenn auch sehr grobes Modell der ungestörten Sprache vor. Die MAM hingegen verwendet ein wesentlich komplexeres Modell der Sprache und vermag so viel besser aus den gestörten Mustern wieder ungestörte Sprachmuster zu rekonstruieren.

Die Vorteile gegenüber dem direkten Einsatz der PMC zur Anpassung des akustischen Modells des Erkenners wurden schon weiter oben angesprochen, daher werden abschließend die Stärken und Besonderheiten der MAM in Bezug auf Störungen durch additives Rauschen nochmals zusammengefasst:

- Geräuschreduktion für komplette Merkmalsvektoren.
- Varianz des Rauschens wird berücksichtigt.
- Kaum freie Parameter.
- Unabhängiger Merkmalsraum (daher auch PMC gut einsetzbar).
- Effiziente Umsetzbarkeit durch Verwendung eines Hilfsmodells.
- Implizite Rekonstruktion fehlender Merkmale.

Vergleich der Sprechernormierung

Die Sprechernormierung über das Likelihood-Kriterium konnte in das neue Verfahren integriert werden, was gegenüber der VTLN mit dem akustischen Modell des Erkenners folgende Vorteile mit sich bringt:

- An verrauschte Daten angepasstes Modell.
- Kein rechenintensiver Erkennungslauf zur Bestimmung einer ersten Hypothese, um eine Zuordnung der Muster zu den Phonemmodellen zu finden.

Während sich der erste Punkt positiv auf die Schätzung der Verzerrungsfaktoren auswirkt, zielt der zweite auf die Erkennungsgeschwindigkeit und bedingt dadurch Einbußen in der Modellierungsgenauigkeit. Eine zuverlässigere Schätzung konnte über eine längere Zeitdauer erreicht werden, womit aber, wie auch schon bei der rechenintensiveren ML-VTLN mit dem Erkennermodell, nur für manche Kategorien Verbesserungen erzielt wurden.

Eigenschaften und Eignung des Verfahrens

Den größten Anteil an der Verbesserung des Basiserkenners hat die neue Art der Geräuschreduktion in Verbindung mit einer abgestimmten Kanalkompensation. Neben einer drastischen Fehlerreduktion bis teilweise über 60 % ergibt sich durch die Rekonstruktion der ungestörten Merkmale auch eine merklich schnellere Erkennungsgeschwindigkeit als mit dem Basiserkenner (mit den hier verwendeten Einstellungen betrug die benötigte Erkennungszeit inklusive MAM zum Teil nur die Hälfte). Durch die Geräuschreduktion reduziert sich nämlich gleichzeitig die Anzahl der untersuchten, noch als vielversprechend angenommenen Pfade im Suchstrahl (siehe S. 68).

Der zusätzliche Rechenaufwand hält sich in Grenzen. Es müssen 100 Gaußdichten kombiniert und pro Sprachrahmen ausgewertet werden, um danach die Verschiebungsvektoren zu berechnen. Dies kann bereits mit heutiger Rechenleistung gut im Rahmen einer Echtzeiterkennung umgesetzt werden. Zieht man hier einen Vergleich mit Adaptionsverfahren, die das akustische Modell des Erkenners verändern, kommen dort neben der großen Anzahl von Dichten (im vorliegenden Fall 80.000) noch weitere Berechnungen hinzu. Bei der unüberwachten MLLR wird eine Klassenzuordnung benötigt, was einen zusätzlichen, sehr zeitintensiven Erkennungslauf voraussetzt. Bei einer PMC für die Modelle des Erkenners kommen aufgrund der verwendeten Merkmale, die normalerweise komplexer sind als beim Hilfsmodell der MAM, weitere Rechenschritte für jede der Gaußdichten hinzu, deren Anzahl 800 mal größer als die des Hilfsmodells der MAM ist.

Es bleibt daher festzuhalten: Mit der MAM konnte, basierend auf den Trainingsdaten aus der Büroumgebung und einigen einfach zu beschaffenden Geräuschaufnahmen, ein Erkenner realisiert werden, der für die unterschiedlichsten Bedingungen gute Ergebnisse erzielt, die sonst nur durch jeweils eines der Referenzsysteme erreicht werden. Damit ist eine Alternative zu der aufwändigen Sammlung und Verschriftung von Daten aus einer neuen Einsatzumgebung gefunden. Dadurch dass die Störungen nicht nur im Training mitgelernt werden, sondern während der Erkennung kompensiert werden, kann man davon ausgehen, dass dieser Erkenner sich auch in völlig neuen Umgebungen robust verhält. Durch die hier verfolgte harte

Randbedingung, die Kompensation nur mithilfe einer einzelnen Äußerung durchzuführen, wird eine Robustheit für schnell wechselnde Umgebungen gewährleistet.

Damit ist ein effektives und zugleich effizientes Verfahren gefunden, mit dem es möglich ist, basierend auf bestehenden Spracherkennern für spezielle rauscharme Umgebungen bzw. mit Sprachdaten aus dieser Umgebung, Anwendungen wie mobile, sprachgesteuerte Informationssysteme besser in die Praxis umzusetzen.

10 Zusammenfassung und Ausblick

Problematik

Automatische Erkennungssysteme für kontinuierliche Sprache arbeiten in Laborumgebung bereits mit hoher Zuverlässigkeit. Beim praktischen Einsatz in mobilen Informationssystemen, wie z.B. im Auto oder bei einem tragbaren Gerät, führen die andersartigen und ständig wechselnden Bedingungen jedoch zu drastischen Leistungseinbußen. Basierend auf einem für Laborumgebung optimierten Erkenner (Abschnitt 4.2) für spontansprachliche Anfragen an ein Navigations- und Informationssystem wurden in dieser Arbeit verschiedene Ansätze zur Steigerung der Robustheit untersucht. Die Experimente wurden nicht wie oft üblich auf simulierten, sondern auf realen, in fahrenden Personenkraftwagen aufgenommenen Sprachdaten durchgeführt. Diese exemplarische Einsatzumgebung wurde gewählt, da sie anspruchsvolle und wechselnde akustische Bedingungen bietet und ein großes wirtschaftliches Interesse an ihrer Beherrschung besteht. Dabei kommen viele Störfaktoren zusammen, die sich gegenseitig beeinflussen, so dass derzeitig eingesetzte Verfahren, die meist nur eine Art Störung berücksichtigen, nicht die gewünschten Verbesserungen erzielen.

Kontinuierlich gesprochene Sprachaufnahmen aus der Autoumgebung standen bislang noch nicht in ausreichender Menge zur Verfügung, da sich die Arbeiten auf diesem Gebiet auf die Erkennung von Einzelwörtern oder Ziffern- und Buchstabensequenzen konzentrieren. Daher wurde als Grundlage für die hier beschriebenen Untersuchungen eine größere Menge Sprachmaterial in der Autoumgebung gesammelt (Abschnitt 4.1). Spontansprachliche Navigationsanfragen in verschiedenen Fahrsituationen dienten dabei zur Beurteilung der Erkennungsleistung. Daneben wurden auch kontinuierliche Sprachdaten zum Trainieren des akustischen Modells eines speziellen Erkenners für die Autoumgebung gesammelt. Ein solches Vorgehen ist jedoch sehr aufwändig und teuer, wird also nicht für alle Anwendungsumgebungen möglich sein.

Es besteht daher ein starkes Interesse auf bereits vorhandene Trainingsdaten aus geräuscharmer Laborumgebung zurückzugreifen. Die darin enthaltenen Muster für verschiedene bedeutungsunterscheidende Sprachlaute sind bei einer homogenen Datenerhebung relativ gut unterscheidbar. Um beliebigen Sprechern die Verwendung des Spracherkenners zu ermöglichen, werden Trainingsmuster von vielen unterschiedlichen Sprechern verwendet. Dies erhöht jedoch die Varianz der Lautmuster und reduziert die Erkennungsleistung im Vergleich zu sprecherabhängigen Modellen. In einer sprachgesteuerten mobilen Anwendung weichen die Sprachmuster außerdem durch veränderten Übertragungskanal und situationsabhängiges Hintergrundrauschen von den Trainingsbedingungen ab („*Mismatch*"). Daher wurden in dieser Arbeit Verfahren betrachtet, die den Zusammenhang zwischen Sprachmustern und akustischem Modell des Erkenners betreffen. Nach einer Einteilung der Verfahren (Kapitel 5) wurden vor allem Ansätze untersucht, die die Störeinflüsse reduzieren

bzw. normieren oder das akustische Modell des Erkenners an die neuen Umgebungsbedingungen anpassen.

Verfahren und Ergebnisse

Das Ausgangssystem, basierend auf dem *Janus Recognition ToolKit* (JRTk) und über mehrere Jahre von der Karlsruher Forschungsgruppe für Laborumgebung optimiert, erzielte für spontansprachliche Navigationsanfragen, die über ein Nahbesprechungsmikrofon in geräuscharmer Umgebung aufgenommen wurden, eine Wortakkuratheit von 86,9 %.

Neben diesem mit Labordaten trainierten Spracherkenner („Labor") wurden zunächst zwei weitere Spracherkennungssysteme erstellt, die Referenzwerte für die Erkennungsleistung in der Autoumgebung lieferten. Die akustischen Modelle beider Systeme wurden mit speziellen Daten neu trainiert, um sie an die neue Umgebung anzupassen. Im ersten Fall wurden dazu die gesammelten Autosprachdaten verwendet („Auto") und im zweiten Fall Labordaten, denen Geräuschaufnahmen aus dem Auto zugemischt wurden („Sim"). Diese Systeme liefern für laborähnliche Testbedingungen (Motor aus) zwar schlechtere Erkennungsraten, schneiden aber unter dem Einfluss von typischen Fahrzeuggeräuschen wesentlich besser ab (siehe Tabelle).

Worterkennungsraten in %	Testbedingung:		
System:	Motor aus	Motor an, 0 km/h	Motor an, 100 km/h
Labor	**86,1**	**78,7**	**38,7**
Sim	**82,8**	**84,0**	**69,4**
Auto	**69,9**	**82,0**	**79,1**

Besonders die Beschaffung und Verschriftung der Spachdaten aus einer speziellen Umgebung (wie für „Auto") ist jedoch sehr aufwändig und verbessert die Leistung nur für gleichartige Bedingungen. Im Weiteren wurden deshalb Methoden untersucht, die nur basierend auf einer Äußerung die Modelle des Erkenners adaptieren, die Muster bezüglich des Sprechers und des Kanals normieren oder das Rauschen reduzieren. Dabei wurde auch auf die effiziente Umsetzbarkeit und die Wechselwirkung zwischen den betrachteten Störeinflüssen geachtet.

Die wichtigsten Beiträge dieser Arbeit sind:

- *ein effektiver, aber auch aufwändiger Adaptionsmechanismus*: Durch die Simulation von Adaptionsdaten mit der hier entwickelten OSTD (Online Simulated Training Data), die die augenblickliche Umgebungssituation einer zu erkennenden Aufnahme wiederspiegeln, konnten die Modelle des Erkenners wesentlich effektiver auf die Situation angepasst werden, als dies mit einer unüberwachten MLLR-Adaption möglich war (→ Kapitel 6, S. 99).

- *eine Sprechernormierung, die mit einer geringen Menge Sprachdaten eines neuen Sprechers auskommt*: Die Sprechernormierung, basierend auf der Vokaltraktlängennormierung (VTLN), wurde für den Einsatz in der Spracherkennung untersucht und verbessert. Durch ein neues Optimierungskriterium (LD-VTLN) wurde die Konvergenzgeschwindigkeit der im Training iterativ zu bestimmenden Normierungsparameter erhöht und die Erkennungsleistung gesteigert (→ Kapitel 7, S. 119).

- *eine Verbesserung der Kanalkompensation für verrauschte Daten*: Die Kanalkompensation wurde für spontansprachliche Äußerungen sowie vor allem für verrauschte Aufnahmen stark verbessert (kontinuierliche 2DCMS, → Kapitel 8, S. 151).

- *ein neues, modellbasiertes Verfahren zur Geräuschreduktion*: Basierend auf einem neu entwickelten Verfahren (MAM) zur Rekonstruktion gestörter Merkmale wurde ein Vorgehen entwickelt, das alle wesentlichen Störeinflüsse berücksichtigt. So konnte die Erkennungsleistung im Fahrzeug eines auf Labordaten trainierten Erkenners sogar über die Werte des „Autospracherkenners" gehoben werden (→ Kapitel 9, S. 156).

Die folgende Tabelle fasst die wichtigsten Ergebnisse im Vergleich mit den Referenzsystemen für drei Testkategorien zusammen:

Worterkennungsraten in %	Testbedingung:		
System:	Motor aus	Motor an, 0 km/h	Motor an, 100 km/h
Labor	**86,1**	**78,7**	**38,7**
Labor + OSTD	86,6	81,5	62,8
Labor + LD-VTLN	86,4	85,5	34,0
Labor + 2DCMS	86,9	86,3	51,5
Labor + MAM	87,6	89,3	70,1
Sim	**82,8**	**84,0**	**69,4**
Sim + MAM	85,2	86,7	79,9
Auto	**69,9**	**82,0**	**79,1**

Besonders die Modellkombinationsbasierte Akustische Transformation (MAM) stellt ein leistungsfähiges und zugleich schnell arbeitendes Verfahren dar. Unter Verwendung von künstlich verrauschten Labordaten (Sim + MAM) erzielt es Ergebnisse, die für jede Testbedingung auf einem Niveau mit dem jeweils besten Referenzsystem liegt. Die erzielten Gewinne gehen nicht auf Kosten einer erhöhten Rechenzeit, da der Mehraufwand für das Verfahren durch eine schnellere Suche aufgrund der entstörten Muster aufgewogen wird.

Ausblick

Mit den hier entwickelten Verfahren konnte die Leistungsfähigkeit eines kontinuierlichen Spracherkenners für spontansprachliche Anfragen an ein Navigations- und Informationssystem in wechselnder und stark gestörter Einsatzumgebung wesentlich gesteigert werden. Er

erreicht nun annähernd Werte, die zuvor nur in geräuscharmer Laborumgebung erzielt wurden. In einem kurzen Ausblick wird nun auf weitere Verbesserungsmöglichkeiten hingewiesen.

In den vorgestellten Verfahren steckt eventuell noch Potential für eine weitere Steigerung der Erkennungsleistung, indem man sie untereinander oder mit anderen Verfahren kombiniert. So könnte das Laborsystem von einer Anpassung seines akustischen Modells (mittels OSTD oder MLLR) an die durch MAM rekonstruierten Muster profitieren, was jedoch den Rechenaufwand beträchtlich erhöhen würde.

Für die durchgeführten Experimente wurden Informationen über die neue Umgebung nur mittels der Testäußerung gewonnen. Dadurch sollte eine Modellanpassung oder Störkompensation für schnell wechselnde Bedingungen und für völlig neue Einsatzorte gewährleisten bleiben. Der Spracherkenner ist somit nicht nur auf die Autoumgebung angepasst, sondern er kann auch in andersartigen Umgebungen eingesetzt werden (wie z.B. Zug, Flugzeug, belebte Fußgängerzone). Wird jedoch ein mobiles Informationssystem längere Zeit vom selben Sprecher oder in der selben Geräuschkulisse benutzt, könnten mehr Daten zur Schätzung von Sprechereigenschaften oder Geräuschstatistiken herangezogen werden. Dazu müssen dem Spracherkenner Informationen über den Benutzer oder die Einsatzumgebung zur Verfügung gestellt werden.

Die vorliegende Arbeit konzentrierte sich ausschließlich auf die akustische Komponente der Spracherkennung. Bei stärkeren Störungen ist die Information aus dem akustischen Signal weniger zuverlässig. Man sollte daher zusätzliche Quellen ausnutzen oder höher bewerten. Dies trifft beispielsweise auf das (linguistische) Sprachmodell zu, das in diesem Falle stärker eingebracht werden könnte. In der Zukunft könnte die Spracherkennung auch von einer stärkeren Kopplung mit den Komponenten zum Sprachverstehen profitieren.

Literatur

„Gedenke der Quelle, wenn du trinkst."

– Chinesisches Sprichwort -

Abkürzungen für häufig erwähnte Konferenzen:

EUROSPEECH European Conference on Speech-Communication and Technology
ICASSP IEEE International Conference on Acoustic, Speech, and Signal Processing
ICSLP International Conference on Speech and Language Processing

Literaturliste:

[Acero 1990] **Alejandro Acero**: "*Acoustical and Environmental Robustness in Automatic Speech Recognition*", Dissertation, Carnegie Mellon University, Pittsburgh, 1990
auch unter:
Kluwer Academic Publishers, Boston/Dordrecht/London, 1993

[Andreou *et al.* 1994] **A. Andreou, T. Kamm, and J. Cohen:** "*Experiments in Vocal Tract Normalization*", Proceedings of the CAIP Workshop: Frontiers in Speech Recognition II, 1994

[Atal 1974] **B.S. Atal:** "*Effectiveness of Linear Prediction Characteristics of Speech Wave for Automatic Speaker Identification and Verification*", Journal of the Acoustic Society of America, 55, pp.1304-1312, 1974

[Aubert *et al.* 1993] **X. Aubert, R. Haeb-Umbach and H. Ney**: "*Continuous Mixture Densities and Linear Discriminant Analysis for Improved Context-Dependent Acoustic Models*", ICASSP, IEEE, Vol. II, pp. 648-651, Minneapolis, 1993

[Baum 1972] **Leonard E. Baum**: "*An Inequality and Associated Maximization Technique in Statistical Estimation for Probabilistic Functions of Markov Chains*", Inequalities, 3, pp. 1-8, 1972

[Berouti *et al.* 1979] **M. Berouti, R. Schwartz, and J. Makhoul**: "*Enhancement of Speech Corrupted by Acoustic Noise*", ICASSP, IEEE, pp. 208-211, 1979

[Bippus *et al.* 1999] **Rolf Bippus, Alexander Fischer and Volker Stahl:** "*DOMAIN ADAPTATION FOR ROBUST AUTOMATIC SPEECH RECOGNITION IN CAR ENVIRONMENTS*", EUROSPEECH, Vol. 5, ESCA, pp. 1943-1946, Budapest, Sep 1999

[Boll 1979] **Steven F. Boll**: "*Suppression of Acoustic Noise in Speech Using Spectral Subtraction*", IEEE Transactions on Signal Processing, Vol. ASSP-27, No. 2, pp. 113-120, Apr 1979

[Boner 1992] **Andreas Boner:** „*Spracherkennung mit Computer*", AT Verlag, vde-verlag, 1992

[Bronstein und Semendjajew] **I.N. Bronstein und K.A. Semendjajew**: „*Taschenbuch der Mathematik*", Verlag Harry Deutsch, Thun und Frankfurt/Main, z.B. 23. Aufl. 1987

[Cappé 1994] **Olivier Cappé**: "*Elimination of the Musical Noise Phenomenon with the Ephraim and Malah Noise Suppressor*", IEEE Transactions on Speech and Audio Processing, Vol. 2(2), pp. 345-349, April 1994

[Carstens-Behrens 1996] **Sönke Carstens-Behrens:** „*Adaptive Geräuschreduktion für gestörte Sprachsignale durch Filterung im Spektralbereich*", Diplomarbeit, Christian-Albrechts-Universität zu Kiel, Mai 1996

[Chiang 1997] **TungHui Chiang**: "*Speech Recognition in Noise Using On-line HMM Adaptation*", EUROSPEECH, pp. 1155-1158, Rhodes, Sep 1997

[Claes und Compernolle 1996] **Tom Claes and Dirk Van Compernolle**: "*SNR-NORMALIZATION FOR ROBUST SPEECH RECOGNITION*", ICASSP, IEEE, pp. 331-334, Atlanta, 1996

[Compernolle 1989a] **Dirk Van Compernolle**: "*Spectral Estimation Using a Log-Distance Error Criterion Applied to Speech Recognition*", ICASSP, IEEE, pp. 258-261, Glasgow, 1989

[Compernolle 1989b] **Dirk Van Compernolle**: "*Noise Adaptation in a Hidden Markov Model Speech Recognition System*", Computer Speech and Language, 3(2), pp. 151-167, 1989

[Cooke *et al.*] **Martin Cooke, Andrew Morris, and Phil Green**: "*Missing Data Techniques for Robust Speech Recognition*", ICASSP, IEEE, pp. 863-866, Munich, 1997

[Crafa *et al.* 1998] **Stefano Crafa, Luciano Fissore and Claudio Vair:** "*DATA-DRIVEN PMC AND BAYESIAN LEARNING INTEGRATION FOR FAST MODEL ADAPTATION IN NOISY CONDITIONS*", ICSLP, Sydney, Dec 1998

[Crystal 1995] **David Crystal:** „*Die Cambridge Enzyklopädie der Sprache*", Verlag Neue Zürcher Zeitung, Zürich, 1995

[Dempster *et al.* 1977] **A.P. Dempster, N.M. Laird and D.B. Rubin**: "*Maximum Likelihood from Incomplete Data via the EM Algorithm*", Journal of the Royal Statistical Society, 39(1), pp. 1-22, 1977

[Draxler *et al.* 1999] **Christoph Draxler, Robert Grudszus, Stefan Euler and Klaus Bengler:** "*First Experiments of the German SpeechDat-Car Database Collection in Mobile Environments*", EUROSPEECH, Vol. 2, ESCA, pp. 919-912, Budapest, Sep 1999

[Duda und Hart 1973] **Richard O. Duda and Peter E. Hart**: "*Pattern Classification and Scene Analysis*", John Wiley & Sons, New York, 1973

[Dülfer 1993] **Bernd Ulrich Dülfer:** „*Klassifikation und Merkmalsextraktion*", Inaugural-Dissertation, Albert-Ludwig-Universität, Freiburg i. Br., Jan. 1993

[Dupont 1998] **Stephane Dupont:** "*Missing Data Reconstruction for Robust Automatic Speech Recognition in the Framework of Hybrid HMM/ANN Systems*" ,ICSLP, Vol. 4, pp. 1439-1442, Sydney, Dec 1998

[Eide und Gish 1996] **Ellen Eide and Herbert Gish**: "*A Parametric Approach to Vocal Tract Length Normalization*", ICASSP, IEEE, pp. 346-348, Atlanta, 1996

[Ephraim und Malah 1983] **Y. Ephraim and D. Malah**: "*Speech Enhancement Using Optimal Nonlinear Spectral Amplitude Estimation*", ICASSP, IEEE, pp. 1118-1121, Boston, 1983

[Ephraim und Malah 1984] **Y. Ephraim and D. Malah**: "*Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator*", IEEE Transactions on Acoustic, Speech, and Signal Processing, Vol. 33, No. 6, pp. 1109-1121, 1984

[Ephraim und Malah 1985] **Y. Ephraim and D. Malah**: "*Speech Enhancement Using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator*", IEEE Transactions on Acoustic, Speech, and Signal Processing, Vol. 33, No. 2, pp. 443-445, 1985

[Finke *et al.* 1997] **Michael Finke, Petra Geutner, Hermann Hild, Thomas Kemp, Klaus Ries and Martin Westphal**: "*THE KARLSRUHE-VERBMOBIL SPEECH RECOGNITION ENGINE*", ICASSP, IEEE, Munich, 1997

[Fischer und Stahl 1998] **Alexander Fischer and Volker Stahl**: "*SUBWORD UNIT BASED SPEECH RECOGNITION IN CAR ENVIRONMENTS*", ICASSP, IEEE, pp. 257 - 260, Seattle, May 1998

[Fischer und Stahl 1999] **Alexander Fischer and Volker Stahl**: "*DATABASE AND ONLINE ADAPTATION FOR IMPROVED SPEECH RECOGNITION IN CAR ENVIRONMENTS*", ICASSP, IEEE, Phoenix, Mar 1999

[Fritsch und Rogina 1996] **Jürgen Fritsch and Ivica Rogina**: "*The Bucket Box Intersection (BBI) Algorithm for Fast Approximative Evaluation of Diagonal Mixture Gaussians*" ICASSP, Vol. I, IEEE, pp. 837-840, Atlanta, 1996

[Fukunaga 1972] **K. Fukunaga:** "*Introduction to statistical pattern recognition*", 1. Aufl., Academic Press, New York and London, 1972

[Furui 1981] **S. Furui**: "*Cepstral Analysis Technique for Automatic Speaker Verification*", IEEE Transactions on Acoustic, Speech, and Signal Processing, Vol. 29, No. 2, pp. 254-272, 1981

[Furui 1986] **S. Furui**: "*Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum*", IEEE Transactions on Acoustic, Speech, and Signal Processing, Vol. 34, No. 1, pp. 52-59, 1986

[Gales 1995] **Mark John Francis Gales**: "*MODEL-BASED TECHNIQUES FOR NOISE ROBUST SPEECH RECOGNITION*", Dissertation, Gonville and Caius College, University of Cambridge, Sep 1995

[Gales und Woodland 1996] **M.J.F. Gales and P.C. Woodland:** *"Mean and variance adaptation within the MLLR framework"*, Academic Press, Computer Speech and Language, Vol. 10, pp. 249-264, 1996

[Gales und Young 1992] **M.J.F. Gales and S. J. Young**: "*AN IMPROVED APPROACH TO THE HIDDEN MARKOV MODEL DECOMPOSITION OF SPEECH AND NOISE*", ICASSP, Vol. I, IEEE, pp. 233-236, {San Francisco}, Mar 1992

[Gales und Young 1993a] **M.J.F. Gales and S. J. Young**: "*HMM RECOGNITION IN NOISE USING PARALLEL MODEL COMBINATION*", EUROSPEECH, pp. 837-840, Berlin, Sep 1993

[Gales und Young 1993b] **M.J.F. Gales and S. J. Young**: "*PARALLEL MODEL COMBINATION FOR SPEECH RECOGNITION IN NOISE*", Technical Report, Cambridge University Engineering Department, TR 135, Jun 1993

[Gales und Young 1994] **M.J.F. Gales and S. J. Young**: "*ROBUST CONTINUOUS SPEECH RECOGNITION USING PARALLEL MODEL COMBINATION*", Technical Report, Cambridge University Engineering Department, TR 172, Mar 1994

[Gales und Young 1995] **M.J.F. Gales and S. J. Young**: "*A FAST AND FLEXIBLE IMPLEMENTATION OF PARALLEL MODEL COMBINATION*", ICASSP, IEEE, pp. 133-136, Detroit, 1995

[Gauvain und Lee 1994] **Jean-Luc Gauvain and Chin-Hui Lee**: "*Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains*", TransSAP, Vol. 2, No. 2, IEEE, pp. 291-298, Apr 1994

[Gelin und Junqua 1999] **Philippe Gelin and Jean-Claude Junqua**: "*Techniques for Robust Speech Recognition in the Car Environment*", EUROSPEECH, ESCA, Vol. 6, pp. 2483-2486, Budapest, Sep 1999

[Geutner *et al.* 1998] **Petra Geutner, Matthias Denecke, Uwe Meier and Martin Westphal:** "*CONVERSATIONAL SPEECH SYSTEMS FOR ON-BOARD CAR NAVIGATION AND ASSISTANCE*", ICSLP, Sydney, Dec 1998

[Giuliani 1999] **Diego Giuliani**: "An On-line Acoustic Compensation Technique for Robust Speech Recognition", EUROSPEECH, ESCA, Vol. 6, pp. 2487-2490, Budapest, Sep 1999

[Gong 1995] **Yifan Gong**: "*Speech recognition in noisy environment: A survey*", Speech Communication 16, pp. 261-291, 1995

[Gong und Godfrey 1999] **Y. Gong and John J. Godfrey**: "*TRANSFORMING HMMS FOR SPEAKER-INDEPENDENT HANDS-FREE SPEECH RECOGNITION IN THE CAR*", ICASSP, IEEE, Phoenix, Mar 1999

[Gouvea und Stern 1997] **Evandro B. Gouvea and Richard M. Stern**: "*Speaker Normalization through Formant-Based Warping of the Frequency Scale*", EUROSPEECH, pp.1139-1142, Rhodes, Sep 1997

[Gray *et al.* 1976] **R. Gray, A. Buzo, A. Gray, and Y. Matusyama:** "*Distance Measures for Speech Processing*", IEEE Trans. Acoustics, Speech and Signal Processing ASSP-24, pp. 380-391, Oct. 1976

[Gu *et al.* 1991] **H. Gu, C. Tseng and L. Lee**: "*Isolated-Utterance Speech Recognition using Hidden Markov Models with Bounded State Durations*", IEEE Transactions on Signal Processing, Vol. 39, No. 8, pp. 1743-1752, 1991

[Gupta *et al.* 1996] **Sunil K. Gupta, Frank Soong and Raziel Haimi-Cohen**: "*HIGH-ACCURAY CONNECTED DIGIT RECOGNITION FOR MOBILE APPLICATIONS*", ICASSP, IEEE, pp. 57-60, Atlanta, 1996

[Haeb-Umbach und Ney 1992] **R. Haeb-Umbach and H. Ney**: "*Linear Discriminant Analysis for Improved Large Vocabulary Continuous Speech Recognition*", ICASSP, IEEE, Vol. I, pp. 13-16, San Francisco, 1992

[Haiber 1998] **Udo Haiber:** „*Sprecheradaption in einem Spracherkennungssystem mit stochastischer Modellierung*", Shaker Verlag, Aachen, 1998

[Hermansky 1990] **H. Hermansky:** "*Perceptual linear predictive (PLP) analysis for speech*", Journal of the Acoustic Society of America, pp.1738-1752, 1990

[Hermansky *et al.* 1993] **Hynek Hermansky, Nelson Morgan and Hans-Gunter Hirsch**: "*Recognition of Speech in Additive and Convolutional Noise based on RASTA Spectral Processing*", ICASSP, IEEE, Vol. II, pp. 83-86, Minneapolis, 1993

[Hermansky und Morgan 1994] **H. Hermansky and N. Morgan**: "*RASTA Processing of Speech*", IEEE Transactions on Speech and Audio Processing, 2(4), pp. 578-589, 1994

[Hild 1997] **Hermann Hild:** „*Buchstabiererkennung mit neuronalen Netzen in Auskunftssystemen*", Shaker Verlag, Aachen, 1997

[Hirsch und Ehrlicher 1995] **H. Hirsch and C. Ehrlicher**: "*Noise estimation techniques for robust speech recognition*", ICASSP, IEEE, pp. 153-156, Detroit, 1995

[Holmes 1991] **John N. Holmes:** „*Sprachsynthese und Spracherkennung*", R. Oldenbourg Verlag, München, 1991

[Jabloun und Çetin 1999] **Firas Jabloun and A. Enis Çetin**: "*THE TEAGER ENERGY BASED FEATURE PARAMTERS FOR ROBUST SPEECH RECOGNITION IN CAR NOISE*", ICASSP, IEEE, Phoenix, Mar 1999

[Jayant und Noll 1984] **N.S. Jayant and P. Noll:** "*Digital Coding of Waveforms, Principles and Applications to Speech and Video*", Prentice-Hall, New Jersey, 1984

[Junqua 1996] **Jean-Claude Junqua**: "*The influence of acoustics on speech production: A noise-induced stress phenomenon known as the Lombard reflex*", Speech Communication 20, pp. 13-22, 1996

[Junqua *et al.* 1996] **Jean-Claude Junqua, Steven Fincke and Ken Field**: "*Influence of the Speaking Style and the Noise Spectral Tilt on the Lombard Reflex and Automatic Speech Recognition*", ICSLP, Vol. 2, pp. 467-470, Sydney, Dec 1998

[Junqua und Haton 1996] **J-C. Junqua and J.-P. Haton**: "*Robustness in Automatic Speech Recognition*", Kluwer Academic Publishers, 1996

[Kammeyer und Kroschel 1989] **Karl Dirk Kammeyer und Kristian Kroschel:** „*Digitale Signalverarbeitung, Filterung und Spektralanalyse*", Teubner, Stuttgart, 1989

[Kemp *et al.* 1998] **Thomas Kemp, Petra Geutner, Michael Schmidt and Borislaw Tornax, Manfred Weber, Martin Westphal and Alex Waibel:** *"THE INTERACTIVE SYSTEMS LABS VIEW4YOU VIDEO INDEXING SYSTEM"*, ICSLP, Sydney, Dec 1998

[Kim *et al.* 1994] **W.-G. Kim, J.-Y. Yoon and D.H. Youn**: *"HMM with Global Path Constraint in Viterbi Decoding for Isolated Word Recognition"*, ICASSP, IEEE, pp. 605-608, Adelaide, 1994

[Kohler 1977] **K. J. Kohler:** *„Einführung in die Phonetik des Deutschen, Grundlagen der Germanistik"*, Erich Schmidt Verlag, Berlin, 1977

[Korthauer 1999] **Andreas Korthauer**: *"Robust Estimation of the SNR of Noisy Speech Signals for the Quality Evaluation of Speech Databases"*, Proceedings of the Workshop on Robust Methods for Speech Recognition in Adverse Conditions, pp. 123-126, Tampere, May 1999

[Kroschel 1986] **Kristian Kroschel:** *„Statistische Nachrichtentheorie, Erster Teil, Signalerkennung und Parameterschätzung"*, 2. Aufl., Springer-Verlag, Berlin, 1986

[Kroschel 1988] **Kristian Kroschel:** *„Statistische Nachrichtentheorie, Zweiter Teil, Signalschätzung"*, 2. Aufl., Springer-Verlag, Berlin, 1988

[Langmann *et al.* 1998] **Detlev Langmann, Hartmut R. Pfitzinger, Tobias Schneider, Robert Grudszus, Alexander Fischer, Martin Westphal, Torsten Crull and Ute Jekosch**: *"CSDC – The MoTiV Car Speech Data Collection"*, Proceedings of the First International Conference on Language Resources and Evaluation (ICLRE), May 1998

[Laurila 1997] **Kari Laurila**: *"Noise Robust Speech Recognition with State Duration Constraints"*, ICASSP, IEEE, pp. 871-874, Munich, 1997

[Lee und Rose 1996] **Li Lee and Richard C. Rose**: *"Speaker Normalization using Efficient Frequency Warping Procedures"*, ICASSP, IEEE, pp. 353-356, Atlanta, 1996

[Legetter und Woodland 1994] **C.J. Legetter and P.C. Woodland**: *„Speaker Adaptation of HMMs using Linear Regression"*, Technical Report, Cambridge University Engineering Department, TR 181, June 1994

[Legetter und Woodland 1995] **C.J. Legetter and P.C. Woodland:** *"Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models"*, Academic Press, Computer Speech and Language, Vol. 9, pp. 171-185, 1995

[Lincoln *et al.* 1997] **M. Lincoln, S. Cox and S. Ringland**: *"A Fast Method of Speaker Normalisation using Formant Estimation"*, EUROSPEECH, Vol. 4, pp. 2095-2098, Rhodes, Sep 1997

[Lippman und Carlson 1997] **Richard P. Lippmann and Beth A. Carlson**: "*Using Missing Feature Theory to actively select Features for Robust Speech Recognition with Interruptions, Filtering, and Noise*", EUROSPEECH, No. KN-37, Rhodes, Sep 1997

[Lockwood und Boudy 1992] **P. Lockwood and J. Boudy**: "*Experiments with a Nonlinear Spectral Subtractor (NSS), Hidden Markov Models and the projection, for robust speech recognition in cars*", Speech Communication, Vol. 11, pp. 215-228, 1992

[Maier 1994] **Martin Maier:** „*Dimensionalitätsreduktion von Sprachsignalen mit statistischen und neuronalen Methoden*", Diplomarbeit, Universität Karlsruhe, 1994

[Martin 1993] **Rainer Martin**: "*An Efficient Algorithm to Estimate the Instantaneous SNR of Speech Signals*", EUROSPEECH, pp. 1093-1096, Berlin, Sep 1993

[Martin *et al.* 1993] **Franck Martin, Kiyohiro Shikano and Yasuhiro Minami**: "*Recognition of Noisy Speech by Composition of Hidden Markov Models*", EUROSPEECH, pp. 1031-1034, Berlin, Sep 1993

[Martinez *et al.* 1993] **T. Martinez, S. Berkovic, and K. Schulten**: "*Neural Gas Network for Vector Quantization and its Applications to Time Series Prediction*", IEEE Transactions on neural Networks, 4(4), pp. 558-569, Jul 1993

[Matrouf und Gauvain 1997] **Driss Matrouf and Jean-Luc Gauvain**: "*Model Compensation for Noises in Training and Test Data*", ICASSP, IEEE, pp. 831-834, Munich, 1997

[Miller 1989] **James D. Miller:** "*Auditory-perceptual interpretation of the vowel*", Journal of the Acoustic Society of America, 85(5), pp.2114-2134, May 1989

[Moreno *et al.* 1995] **Pedro J. Moreno, Bhiksha Raj, Evandro Gouvea and Richard M. Stern**: "*Multivariate-Gaussian-Based Cepstral Normalization for Robust Speech Recognition*", ICASSP, IEEE, pp. 137-140, Detroit, 1995

[Morgan *et al.* 1997] **N. Morgan, E. Fosler and N. Mirghafori**: "*Speech Recognition Using On-line Estimation of Speaking Rate*", EUROSPEECH, pp. 2079-2082, Rhodes, Sep 1997

[Muthusamy *et al.* 1999] **Yeshwant Muthusamy, Rajeev Agarwal, Yifan Gong and Vishu Viswanathan**: "*SPEECH-ENABLED INFORMATION RETRIEVAL IN THE AUTOMOBILE ENVIRONMENT*", ICASSP, IEEE, Phoenix, Mar 1999

[Nguyen *et al.* 1999] **P. Nguyen, Ph. Gelin, J-C. Junqua and J-T. Chien**: "*N-BEST BASED SUPERVISED AND UNSUPERVISED ADAPTATION FOR NATIVE AND NON-NATIVE SPEAKERS IN CARS*", ICASSP, IEEE, Phoenix, Mar 1999

[Noll 1974] **P. Noll:** "*Adaptive Quantization in Speech Coding Systems*", Proc. Int. Zurich Seminar on Digital Communications, pp. B3.1-B3.6, Zurich, Oct. 1974

[Openshaw und Mason 1994] **J.P. Openshaw and J.S. Mason:** "*On the Limitations of Cepstral Features in Noise*", ICASSP, IEEE, Vol. II, pp. 49-52, Adelaide, Apr 1994

[Oppenheim und Schafer 1975] **A. V. Oppenheim and R. W. Schafer**: "*Digital Signal Processing*", Prentice-Hall, 1975

[Pfau *et al.* 1999] **Thilo Pfau, Robert Faltlhauser und Günther Ruske**: "*Speaker Normalization and Pronunciation Variant Modeling: Helpful Methods for Improving Recognition of Fast Speech*", EUROSPEECH, Vol. 1, ESCA, pp. 229-232, Budapest, Sep 1999

[Porter und Boll 1984] **J.E. Porter and S.F. Boll**: "*Optimal Estimators for Spectral Restoration of Noisy Speech*", ICASSP, IEEE, pp. 18A.2.1., San Diego, 1984

[Pouteau und Arevalo 1998] **Xavier Pouteau and Luis Arevalo**: "*Robust spoken dialogue systems for consumer products: a concrete application*", ICSLP, Vol. 4, Sydney, Dec 1998

[Rabiner 1989] **L.R. Rabiner**: "*A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition*", Proceedings of the IEEE, 77(2), pp. 257-285, Feb 1989 (auch in [Waibel und Lee 1990, S. 267-296])

[Rabiner und Schafer 1978] **Lawrence R. Rabiner and Ronald W. Schafer**: "*Digital Processing of Speech Signals*", Prentice-Hall, Alan V. Oppenheim, Signal Processing Series, New Jersey, 1978

[Raj *et al.* 1999] **Bhiksha Raj, Rita Singh and Richard M. Stern**: "*Inference of Missing Spectrographic Features for Robust Speech Recognition*", ICSLP, Vol. 4, pp. 1491-1494, Sydney, Dec 1998

[Ramalingam *et al.* 1999] **C.S. Ramalingam, Yifan Gong, Lorin P. Netsch, Wallace W. Anderson, John J. Godfrey and Yu-Hung Kao**: "*SPEAKER-DEPENDENT NAME DIALING IN A CAR ENVIRONMENT WITH OUT-OF-VOCABULARY REJECTION*", ICASSP, IEEE, Phoenix, Mar 1999

[Renevey *et al. 1999*] **Philippe Renevey and Andrzej Drygajlo:** "*Missing Feature Theory and Probabilistic Estimation of Clean Speech Components for Robust Speech Recongition*", EUROSPEECH, Vol. 6, ESCA, pp. 2627-2630, Budapest, Sep 1999

[Rogina 1998] **Ivica Rogina**: „*Parameterraumoptimierung für Diktiersysteme mit unbeschränktem Vokabular*", Shaker Verlag, Aachen, 1998

[Ruske 1994] **G. Ruske**: *„Automatische Spracherkennung - Methoden der Klassifikation und Merkmalsextraktion"*, R. Oldenbourg Verlag, München, 2. Auflage, 1994

[Sankar und Lee 1995] **Ananth Sankar and Chin-Hui Lee**: "Robust Speech Recognition based on Stochastic Matching", ICASSP, IEEE, pp. 121-124, Detroit, 1995

[Schless und Class 1997] **Volker Schless and Fritz Class**: *"Adaptive Model Combination for Robust Speech Recognition in Car Environments"*, EUROSPEECH, pp. 1091-1094, Rhodes, Sep 1997

[Schless und Class 1998] **Volker Schless and Fritz Class**: *"SNR-DEPENDENT FLOORING AND NOISE OVERESTIMATION FOR JOINT APPLICATION OF SPECTRAL SUBTRACTION AND MODEL COMBINATION"*, ICSLP, Sydney, Dec 1998

[Schubert 1999] **Kjell Schubert**: *„Grundfrequenzverfolgung und deren Anwendung in der Spracherkennung"*, Diplomarbeit, Universität Karlsruhe, 1999

[Schukat-Talamazzini 1995] **E. G. Schukat-Talamazzini:** *„Automatische Spracherkennung - Statistische Verfahren der Musteranalyse"*, Vieweg Verlag, Braunschweig, 1995

[Schultz 1994] **Tanja Schultz:** *„Akustische Modellierung sprachlicher und nichtsprachlicher Geräusche"*, Studienarbeit, Universität Karlsruhe, April 1994

[Shieh und Chang 1999] **Woei-Chyang Shieh and Sen-Chia Chang**: *"The Dependence of Feature Vectors under Adverse Noise"*, EUROSPEECH, Vol. 5, ESCA, pp. 2395-2398, Budapest, Sep 1999

[Shozakai 1999] **M. Shozakai**: *"SPEECH INTERFACE VLSI FOR CAR APPLICATIONS"*, ICASSP, IEEE, Phoenix, Mar 1999

[Shozakai *et al.* 1998] **M. Shozakai, S. Nakamura and K. Shikano**: *"Robust Speech Recognition in Car Environments"*, ICASSP, IEEE, pp. 269-272, Seattle, May 1998

[Siegler und Stern 1995] **M.A. Siegler and R.M. Stern**: *"On the Effects of Speech Rate in Large Vocabulary Speech Recognition Systems"*, ICASSP, IEEE, pp. 612-615, Detroit, 1995

[Smolders *et al.* 1994] **Johan Smolders, Tom Claes, Gert Sablon and Dirk Van Compernolle:** *"On the Importance of the Microphone Position for Speech Recognition in the Car"*, ICASSP, Vol. I, IEEE, pp. 429-432, Adelaide, Apr 1994

[Song *et al.* 1998] **Myung Gyu Song, Hoi In Jung, Kab-Jong Shim and Hyung Soon Kim**: *"Speech Recognition in Car Noise Environments using Multiple Models according to Noise Masking Levels"*, ICSLP, Vol. 4, pp. 1475-1478, Sydney, Dec 1998

[Suhm 1996] **Bernhard Suhm, Brad Myers and Alex Waibel:** "*Interactive Recovery from Speech Recognition Errors in Speech User Interfaces*", ICSLP, pp. 865-868, Philadelphia, 1996

[Suhm 1998] **Bernhard Suhm:** "*Multimodal interactive error recovery for non-conversational speech user interfaces*", Dissertation, Universität Karlsruhe, Sep 1998

[Tamura und Waibel 1988] **Shin'ichi Tamura and Alex Waibel:** *"NOISE REDUCTION USING CONNECTIONIST MODELS"*, ICASSP, IEEE, No. S12.7, pp. 553-556, 1988

[Trancoso *et al.* 1999] **Isabel Trancoso, Ceu Viana, Isabel Mascarenhas and Carlos Teixeira**: "*On deriving rules for nativised pronunciation in navigation queries*", EUROSPEECH, Vol. I, ESCA, pp. 195-198, Budapest, Sep 1999

[Trompf 1996] **Michael Trompf:** *„Künstliche neuronale Netzwerke zur adaptiven Geräuschreduktion für robuste Spracherkennung"*, Dissertation, Universität Fridericiana Karlsruhe, Stuttgart, 1996

[Van Compernolle 1997] **Dirk Van Compernolle:** *"Speech Recognition in the Car - From Phone Dialing to Car Navigation"*, EUROSPEECH, pp. 2431-2434, Rhodes, Sep 1997

[Varga und Moore 1990] **A.P. Varga and R.K. Moore**: "*Hidden Markov Model Decomposition of Speech and Noise*", ICASSP, IEEE, pp. 845-848, Albuquerque, 1990

[Veth *et al.* 1999] **Johan de Veth, Bert Cranen, Febe de Wet and Louis Boves:** "*Acoustic Pre-Processing for Optimal Effectivity of Missing Feature Theory*", EUROSPEECH, Vol. 1, ESCA, pp. 65-68, Budapest, Sep 1999

[Veth *et al.*] **Johan de Veth, Bert Cranen, and Louis Boves**: "*Acoustic Backing-Off in the Local Distance Computation for Robust Automatic Speech Recognition*", ICSLP, Vol. 4, pp. 1427-1430, Sydney, Dec 1998

[Veth und Boves 1996] **Johan de Veth and Louis Boves**: "*Comparison of Channel Normalisation Techniques for Automatic Speech Recognition over the Phone*", ICSLP, pp. 2332-2335, Philadelphia, 1996

[Vizinho *et al.* 1999] **A. Vizinho, P. Green, M. Cooke and L. Josifovski:** *"MISSING DATA THEORY, SPECTRAL SUBTRACTION AND SIGNAL-TO-NOISE ESTIMATION FOR ROBUST ASR: AN INTEGRATED STUDY"*, EUROSPEECH, Vol. 5, ESCA, pp. 2407-2410, Budapest, Sep 1999

[Waibel und Lee] **A. Waibel and K.-F. Lee**: "*Readings in Speech Recognition*", Morgan Kaufmann, San Mateo, 1990

[Wakita 1977] **H. Wakita:** "*Normalization of Vowels by Vocal-Tract Length and Its Application to Vowel Identification*", IEEE Transactions on Speech and Audio Processing, 25(2), pp. 183-192, 1977

[Ward 1989] **W. Ward:** "*Modelling Non-verbal Sounds for Speech Recognition*", Proceedings of Speech and Natural Language Workshop, pp 47-50, Oct. 1989

[Wegmann *et al.* 1996] **Steven Wegmann, Don McAllaster, Jeremy Orloff and Barbara Peskin**: "*Speaker Normalization on Conversational Telephone Speech*", ICASSP, IEEE, pp. 339-341, Atlanta, 1996

[Welling *et al.* 1999] **L. Welling, S. Kanthak and H. Ney**: "*IMPROVED METHODS FOR VOCAL TRACT NORMALIZATION*", ICASSP, IEEE, Phoenix, Mar 1999

[Westphal 1997] **Martin Westphal**: "*THE USE OF CEPSTRAL MEANS IN CONVERSATIONAL SPEECH RECOGNITION*", EUROSPEECH, Rhodes, Sep 1997

[Westphal 1998] **Martin Westphal**: „*Datensammlung kontinuierlicher Sprache im Auto*", Dokumentation zur Datensammlung CCDC-98, ILKD, Universität Karlsruhe, August 1998

[Westphal *et al.* 1998] **Martin Westphal, Alex Waibel and Tanja Schultz:** "*LINEAR DISCRIMINANT - A NEW METHOD FOR SPEAKER NORMALIZATION*", ICSLP, Sydney, Dec 1998

[Westphal und Waibel 1999] **Martin Westphal and Alex Waibel:** "*TOWARDS SPONTANEOUS SPEECH RECOGNITION FOR ON-BOARD CAR NAVIGATION AND INFORMATION SYSTEMS*", EUROSPEECH, ESCA, Budapest, Sep 1999

[Wittmann *et al.* 1993] **M. Wittmann, O. Schmidbauer and A. Aktas**: "*Online Channel Compensation for Robust Speech Recognition*", EUROSPEECH, pp. 1251-1254, Berlin, Sep 1993

[Woodland *et al.* 1996] **P.C. Woodland, M.J.F. Gales and D. Pye**: "*IMPROVING ENVIRONMENTAL ROBUSTNESS IN LARGE VOCABULARY SPEECH RECOGNITION*", ICASSP, IEEE, pp. 65-68, Atlanta, 1996

[Woszczyna 1998] **Monika Woszczyna**: "*Fast Speaker Independent Large Vocabulary Continuous Speech Recognition*", Dissertation, Universität Karlsruhe, 1998

[Yang und Haavisto 1995] **Ruikang Yang and Petri Haavisto**: "*Noise Compensation for Speech Recognition in Car Noise Environments*", ICASSP, IEEE, pp. 433-436, Detroit, 1995

[Yang und Haavisto 1996] **Ruikang Yang and Petri Haavisto**: "*AN IMPROVED NOISE COMPENSATION ALGORITHM FOR SPEECH RECOGNITION IN NOISE*", ICASSP, IEEE, pp. 49-52, Atlanta, car data, 1996

[Zhan *et al.* 1997] **Puming Zhan, Martin Westphal, Michael Finke and Alex Waibel**: "*SPEAKER NORMALIZATION AND SPEAKER ADAPTATION - A COMBINATION FOR CONVERSATIONAL SPEECH RECOGNITION*", EUROSPEECH, pp. 2087-2090, Rhodes, Sep 1997

[Zhan und Westphal 1997] **Puming Zhan and Martin Westphal**: "*SPEAKER NORMALIZATION BASED ON FREQUENCY WARPING*", ICASSP, IEEE, Munich, 1997

Index

M

N

O

P

Q

R

S

T

Ü

V

W

Z