UNIVERSITÄT PADERBORN
Die Universität der Informationsgesellschaft

FAKULTÄT FÜR
ELEKTROTECHNIK,
INFORMATIK UND
MATHEMATIK

Modellbasierte Merkmalsverbesserung zur robusten automatischen Spracherkennung in Gegenwart von Nachhall und Hintergrundstörungen

Von der Fakultät für Elektrotechnik, Informatik und Mathematik
der Universität Paderborn

zur Erlangung des akademischen Grades

Doktor der Ingenieurwissenschaften (Dr.-Ing.)

genehmigte Dissertation
von

Dipl.-Math. Alexander Krüger

Erster Gutachter: Prof. Dr.-Ing. Reinhold Häb-Umbach
Zweiter Gutachter: Prof. Dr.-Ing. Klaus Meerkötter

Tag der mündlichen Prüfung: 16. Dezember 2011

Paderborn 2011

Diss. EIM-E/282

Danksagung

Die vorliegende Arbeit entstand während meiner Tätigkeit im Fachgebiet Nachrichtentechnik der Universität Paderborn. Während der ersten drei Jahre gehörte ich dabei dem Graduiertenkolleg des Paderborn Institute for Scientific Computation (PaSCo) als Mitglied an, für dessen wissenschaftliche und finanzielle Förderung in Form eines Stipendiums ich mich hiermit herzlich bedanke.

Mein besonderer Dank gilt dem Leiter des Fachgebiets Nachrichtentechnik, Herrn Prof. Dr.-Ing. Reinhold Häb-Umbach, für eine angenehme Arbeitsatmosphäre sowie eine hervorragende Betreuung durch zahlreiche fachliche Ratschläge und Diskussionen, die wesentlich zum Erfolg der Arbeit beigetragen haben. Weiterhin danke ich Herrn Prof. Dr.-Ing. Klaus Meerkötter für die Übernahme des Korreferates dieser Arbeit und ebenfalls für viele fachliche Gespräche, die sich stets als positive Denkanregungen auch im Hinblick auf diese Arbeit erwiesen.

Allen meinen Arbeitskollegen im Fachgebiet Nachrichtentechnik danke ich für ihre stetige Hilsbereitschaft und eine freundschaftliche Arbeitsatmosphäre. Einen besonderen Dank möchte ich in diesem Zusammenhang Herrn Dipl.-Ing. Volker Leutnant, Herrn Dipl.-Ing. Dang Hai Tran Vu, Herrn Dr.-Ing. Maik Bevermeier sowie Herrn Dr.-Ing. Jörg Schmalenströer aussprechen. In zahlreichen fachlichen Diskussionen mit ihnen sind viele wertvolle Ideen entstanden, die in diese Arbeit eingeflossen sind.

Herrn Dipl.-Ing. Volker Leutnant sowie meinem Bruder, Herrn Waldemar Krüger, danke ich für ein sorgfältiges Korrekturlesen dieser Dissertation und für das Anbringen von konstruktiver Kritik.

Schließlich gilt ein großer Dank meinen Eltern, die mich während der gesamten Zeit bedingungslos unterstützt haben. Dadurch hat sich für mich überhaupt erst die Möglichkeit für eine wissenschaftliche Laufbahn eröffnet.

Inhaltsverzeichnis

"Essentially, all models are wrong, but some are useful."

George E. P. Box [BD86]

1. Einleitung

Die automatische Spracherkennung bezeichnet den Prozess der Konversion eines akustischen Signals in eine Menge von Wörtern [CZ98] und kann unter anderem zum Informationsaustausch zwischen dem Menschen und dem Computer genutzt werden. Obwohl sich für diesen Zweck theoretisch auch visuelle oder taktile Kommunikationskanäle eignen, bietet die gesprochene Kommunikation einige entscheidende Vorteile. Zum einen lässt sich damit beispielsweise gegenüber der Eingabe von Wörtern und Zeichen über eine Tastatur oder in handschriftlicher Form mittels eines graphischen Tabletts eine deutlich höhere Datenrate erzielen [ST95]. Zum anderen erfordert die gesprochene Kommunikation in der Regel keine langwierigen Schulungsphasen, wie sie zum Beispiel zum Erreichen einer vernünftigen Schreibgeschwindigkeit mit der Tastatur in der Regel notwendig sind. Das liegt daran, dass die Sprache wohl das wichtigste und natürlichste Mittel der zwischenmenschlichen Kommunikation darstellt und daher von den meisten Menschen in gewissem, ausreichenden Maße beherrscht wird. Ein weiterer wesentlicher Vorteil besteht in dem Erhalt des vollen Funktionsumfanges unter bestimmten, erschwerenden Umständen, wie zum Beispiel bei Dunkelheit oder extremen Bewegungseinschränkungen [ST95]. Ein unbezweifelbarer Nachteil liegt in der gewöhnlich deutlich geringeren Erkennungsleistung im Vergleich zur Eingabe über die Tastatur, wo von einer nahezu hundertprozentigen Erkennung auszugehen ist [ST95].

Eine zentrale Schwierigkeit der automatischen Spracherkennung ist die Variabilität eines Sprachsignals, welches eine hochredundante Codierung einer zu übermittelnden Nachricht darstellt. Damit ist gemeint, dass dieselbe Nachricht prinzipiell auf viele Arten ausgesprochen werden kann, die sich unter anderem in der Sprechweise und in den individuellen und habituellen Sprechermerkmalen unterscheiden. Insbesondere treten auch kontextuelle Aussprachevariationen auf [ST95]. Zur Berücksichtigung der Variabilität basieren aktuelle Spracherkennungssysteme gewöhnlich auf einer statistischen Beschreibung der Sprache mit Modellen, deren Parameter mit Hilfe von Datenbanken vor der eigentlichen Erkennungsphase geschätzt werden. So besteht ein Spracherkenner im Allgemeinen aus zwei Einheiten, der sogenannten Merkmalsextraktion und der Decodierung. Bei der Merkmalsextraktion wird versucht, aus dem Sprachsignal den redundanten Anteil der Information zu entfernen, so dass anschließend bei der Decodierung mit dem relevanten Anteil der Information die eigentliche Suche nach der zugrunde liegenden Wortsequenz stattfinden kann. Mit diesem Ansatz wurden in der Vergangenheit enorme Fortschritte in der automatischen Erkennung von sowohl einzelnen Wörtern als auch von kontinuierlich gesprochener Sprache erzielt [Ata95]. Diese Fortschritte motivierten die Entwicklung einer breiten Palette von kommerziellen Anwendungen und Produkten, welche in den Bereichen der Datenerfassung, der Steuerung von Systemen oder Geräten sowie der automatischen Informationsgewinnung liegen [ST95]. Die potentiellen Anwendungsgebiete umfassen dabei unter anderem Haushalt, Büro, Industrie, Medizin und Militär.

Trotz der enormen Fortschritte in der automatischen Spracherkennung bleibt festzustellen, dass die Leistungsfähigkeit eines Menschen nur unter kontrollierten Aufnahmebedingungen annähernd erreicht wird [Ata95, MS95]. In realistischen Anwendungen können die Aufnahmebedingungen jedoch aufgrund der Verwendung von unterschiedlichen Mikrophonen sowie der etwaigen Präsenz von akustischen Störquellen drastisch variieren. Ein in diesem Zusammenhang für diese Dissertation wesentlicher Aspekt ist die Variation durch die Verwendung von Freisprechsystemen, wobei Fernfeld- an Stelle von Nahbereichsmikrophonen zum Einsatz kommen. Solche Systeme sind für bestimmte Anwendungen entweder unerlässlich oder können zur Steigerung des Komforts und der Sicherheit beitragen [SK08]. Man denke dabei beispielsweise an die kabellose Bedienung von medizinischen diagnostischen Geräten durch einen Chirurgen während einer Operation oder aber auch an die Bedienung eines Fernsehers mittels Sprachsteuerung durch einen Konsumenten [SK08].

Nun führt der erhöhte Abstand des Sprechers zum Mikrophon, der im Bereich von etwa einem bis mehreren Metern liegt, einerseits dazu, dass sich die akustischen Signale eventuell vorhandener Störquellen deutlich stärker bemerkbar machen. Andererseits wird das akustische Signal des gewünschten Sprechers an Oberflächen von Wänden und Gegenständen reflektiert und erfährt dadurch eine sogenannte Mehrwegeausbreitung. Das aufgenommene Signal beinhaltet dann neben dem gewünschten, gedämpften Sprachsignal dessen unterschiedlich zeitlich verzögerte und gedämpfte Versionen, welche in ihrer Gesamtheit als Nachhall bezeichnet werden. Während der Einfluss der additiven Hintergrundstörungen vom Verhältnis zwischen der Energie des Sprach- und des Störsignals abhängt und prinzipiell durch ein bewusstes lauteres Sprechen verringert werden kann, trifft dies für den Nachhall nicht zu, da er eine Faltungsstörung darstellt.

Die soeben beschriebene Variabilität des Sprachsignals, die mit der Verwendung von Freisprecheinrichtungen einhergeht, spiegelt sich erwartungsgemäß in dessen statistischen Eigenschaften wider. Unterscheiden sich diese von denen, welche zum Zeitpunkt des Trainings vorlagen, muss aufgrund dieser Diskrepanz mit beträchtlichen Einbußen in der Leistungsfähigkeit des Spracherkennungssystems gerechnet werden.

Gewöhnlich ist der Mensch jedoch nicht gewillt, zugunsten eines erhöhten Kommunikationskomforts durch Freisprechsysteme eine geringere Erkennungsleistung hinzunehmen. Insbesondere werden diesbezüglich an einen automatischen Spracherkenner oft dieselben Erwartungen wie an einen Menschen gestellt, der nur in geringem Maße empfindlich gegenüber Nachhall sowie Hintergrundstörungen ist. Zur Erfüllung dieser Erwartungen besteht ein hohes Interesse an Methoden zur robusten Spracherkennung.

Während die Forschung im Bereich der Robustheit gegenüber additiven Hintergrundstörungen bereits einige Jahrzehnte andauert [SK08], reichen die ersten Bemühungen um die Robustheit gegenüber dem Nachhall etwa in das Ende der Neunziger Jahre zurück. Trotz intensiver Forschung reicht die Leistungsfähigkeit von automatischen Spracherkennern im Freisprechbetrieb bei Weitem nicht an die eines Menschen heran.

Im Rahmen dieser Arbeit wird nun ein neuartiges Verfahren zur modellbasierten Verbesserung akustischer Merkmale zur robusten Spracherkennung in Gegenwart von Nachhall sowie Hintergrundstörungen vorgestellt, wobei der Fokus deutlich auf der Behandlung des Nachhalls liegt. Der einleitende Teil dieser Dissertation ist wie folgt aufgebaut. Zunächst werden in Kap. 2 die Grundlagen der statistisch motivierten Spracherkennung vorgestellt. Dabei werden die beiden weiter oben erwähnten Bestandteile eines Spracherkennungssystems, nämlich die Merkmalsextraktion und die Decodierung, ausführlich beschrieben. Anschließend wird

in Kap. 3 eine Übersicht über die in der Literatur bisher existenten Methoden zur hallrobusten Spracherkennung gegeben. In Kap. 4 werden dann die wissenschaftlichen Ziele der Arbeit formuliert. Darin findet sich auch der detaillierte Aufbau der weiteren Arbeit.

2. Grundlagen zur automatischen Spracherkennung

Die in der heutigen Literatur existenten Ansätze zur automatischen Spracherkennung können grob in drei Kategorien unterteilt werden [RJ93]:

1. Akustisch-phonetische Ansätze
2. Mustererkennungsansätze
3. Ansätze basierend auf künstlicher Intelligenz.

Die akustisch-phonetischen Ansätze gehen von der Annahme aus, dass sich ein Sprachsignal aus einer Folge von endlichen, unverwechselbaren phonetischen Einheiten zusammensetzt. Dementsprechend besteht der Prozess der Erkennung im Wesentlichen aus einer sinnvollen Segmentierung des Signals, einer anschließenden Zuordnung der Segmente zu Phonemen und einer darauf aufbauenden Bestimmung der zugrunde liegenden Wortsequenz gemäß vorgegebenen syntaktischen und semantischen Regeln sowie einer durch ein Lexikon gegebenen Menge an gültigen Wörtern.

Die Mustererkennungsansätze basieren auf der Vorstellung, dass die Aussprache von Wörtern oder Wortuntereinheiten bestimmte (nicht unbedingt an die Phonetik gebundene) Muster im Sprachsignal hervorruft, die in einer vorhergehenden Trainingsphase gelernt werden müssen. Die eigentliche Erkennung wird dann als Klassifikationsaufgabe aufgefasst, die auf einem Vergleich zwischen den trainierten und den zu klassifizierenden Mustern beruht.

Die auf künstlicher Intelligenz basierenden Ansätze stellen im Prinzip eine Kombination beider vorhergehender Ansätze dar. Sie versuchen den Vorgang der Erkennung derart zu gestalten, wie eine menschliche Person ihre Intelligenz anwenden würde, um das Sprachsignal zu analysieren und eine abschließende Entscheidung über die vermeintliche Wortsequenz zu fällen. Dabei werden für einzelne Teilaufgaben der Erkennung eine große Anzahl von verschiedenen Informationsquellen herangezogen. Beispielsweise zählen dazu Verfahren, welche bereits für die Segmentierung des Sprachsignals abgesehen von dem rein akustisch-phonetischen unter anderem auch lexikalisches, syntaktisches und semantisches Wissen nutzen. Eine große Untergruppe dieser Kategorie bilden die sogenannten neuronalen Netze, mit Hilfe derer (nichtlineare) Zusammenhänge zwischen unterschiedlichen Kontextinformationen gelernt werden können.

Der Fokus dieser Arbeit richtet sich auf die **H**idden-**M**ARKOV-**M**odell (HMM)-basierte Spracherkennung, welche zur Klasse der Mustererkennungsansätze gehört und die heute die am weitesten verbreitete Methode darstellt. Die Grundlage dieser Art der Spracherkennung bildet die Annahme, dass das Sprachsignal als eine Realisierung eines parametrischen Zufallsprozesses charakterisiert werden kann. Sie beruht auf der Tatsache, dass die einem Wort

zugrunde liegende Sequenz von akustischen Merkmalen in der Regel von unterschiedlichen Einflussfaktoren wie dem Alter, dem Geschlecht und dem Gemütszustand des Sprechers, der Sprachgeschwindigkeit, der Intonation usw. abhängt und somit Variationen erfährt.

Der prinzipielle Aufbau eines derartigen statistischen Spracherkennungssystems ist in Abb. 2.1 dargestellt [You08].

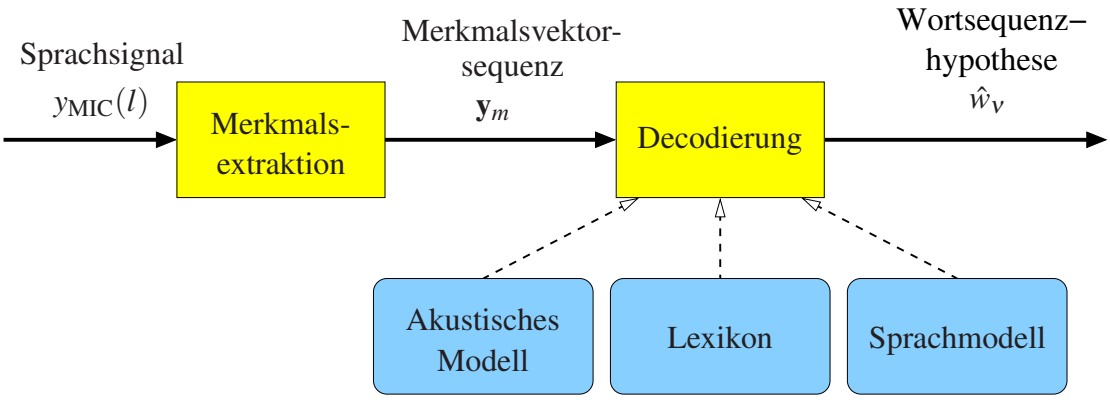


Abbildung 2.1.: *Prinzipieller Aufbau eines statistischen Spracherkennungssystems.*

Ein solches System gliedert sich grob in zwei Untereinheiten. Bei der sogenannten Merkmalsextraktion findet eine akustische Vorverarbeitung statt, bei der aus einem zeitdiskreten Sprachsignal $y_{\mathrm{MIC}}(l)$ akustische Merkmale berechnet werden, welche die für die Erkennung relevante Information tragen. Dabei wird davon ausgegangen, dass das entsprechende zeitkontinuierliche Sprachsignal zuvor bereits sinnvoll tiefpassgefiltert und einer **A**nalog-**D**igital-**U**mwandlung (ADU) unterzogen wurde.

Anschließend erfolgt eine Decodierung der extrahierten Merkmalsvektorsequenz $\mathbf{y}_m$ in eine hypothetische Wortsequenz $\hat{w}_V$. Für die Decodierung werden gewöhnlich die drei folgenden statistischen Informationsquellen verwendet. Das *akustische Modell* beschreibt die akustische Realisierung von Wörtern oder Wortuntereinheiten wie Triphonen, wobei die Menge an zulässigen Wörtern sowie deren mögliche Zusammensetzung aus Wortuntereinheiten durch das *Lexikon* spezifiziert wird. Das *Sprachmodell* beschreibt die Auftrittswahrscheinlichkeit von bestimmten Wörtern oder Wortfolgen. Die Parameter dieser drei Informationsquellen werden vor der eigentlichen Erkennung mit Hilfe von Trainingsdaten geschätzt.

In den beiden folgenden Abschnitten werden die beiden Untereinheiten des aufgeführten Spracherkennungssystems, Merkmalsextraktion und Decodierung, detaillierter beschrieben, da sie die Grundlage für das weitere Verständnis der Arbeit bilden.

2.1. Merkmalsextraktion

Die Merkmalsextraktion verfolgt das Ziel einer parametrischen Repräsentation der akustischen Daten [DM80]. Im Hinblick auf die folgende automatische Spracherkennung erfolgt dabei eine Informationskompression derart, dass jegliche für die phonetische Analyse irrelevanten Aspekte entfernt werden und dass bestenfalls nur diejenige Information verbleibt, welche in hohem Maße dazu beiträgt, phonetische Unterschiede zu detektieren. Die in der Literatur am weitesten verbreiteten Methoden zur Merkmalsextraktion basieren entweder auf

einer spektralen Filterbankanalyse oder einer linearen Prädiktionskodierung (engl. ***L**inear **P**redictive **C**oding* (*LPC*)) [RJ93, GPAF04, YEG+06].

An dieser Stelle soll nun eine detaillierte Beschreibung der sogenannten MEL-Frequenz-Cepstrum-Koeffizienten (engl. ***M**el **F**requency **C**epstral **C**oefficient*s (*MFCC*s)) erfolgen, welche in die erste Kategorie eingeordnet werden können. Ihre Berechnung basiert auf einer Kurzzeit-Spektralanalyse und orientiert sich stark an der menschlichen Gehörwahrnehmung. Sie wurden ursprünglich von Davis und Mermelstein [DM80] eingeführt und sind heutzutage durch das europäische Institut für Telekommunikationsnormen (engl. ***E**uropean **T**elecommunications **S**tandards **I**nstitute* (*ETSI*)) standardisiert [ETSb]. Abbildung 2.2 zeigt ein Blockschaltbild zur Extraktion von *MFCC*s aus einem zeitdiskreten akustischen Signal $y_{\mathrm{MIC}}(l)$ gemäß einer leichten Abwandlung dieses Standards. Diese Art der Vorverarbeitung ist in der englischsprachigen Literatur unter dem Namen ***S**tandard **F**ront **E**nd* (*SFE*) bekannt.

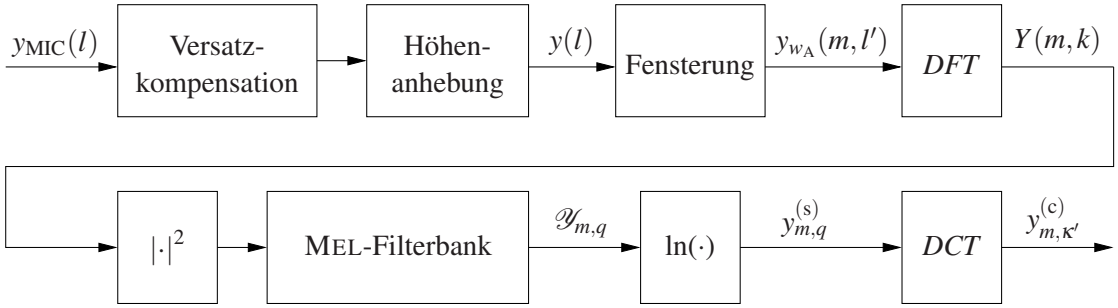


Abbildung 2.2.: *Blockschaltbild zur Extraktion von MFCCs aus einem zeitdiskreten akustischen Signal $y_{\mathrm{MIC}}(l)$ gemäß einer leichten Abwandlung des ETSI-Standards [ETSb]. Die Änderung gegenüber [ETSb] besteht in der Ersetzung des Kurzzeit-Amplitudenspektrums durch das Kurzzeit-Leistungsspektrum zur Vereinfachung der Berechnung.*

Wie bereits weiter oben angesprochen wird davon ausgegangen, dass das entsprechende zeitkontinuierliche Sprachsignal bereits einer Tiefpassfilterung sowie einer ADU unterzogen wurde, wobei die Abtastfrequenz mit f_{A} und die Abtastdauer mit $T_{\mathrm{A}} = \frac{1}{f_{\mathrm{A}}}$ bezeichnet werden soll. Obwohl der *ETSI*-Standard [ETSb] für die drei Abtastfrequenzen 8 kHz, 11 kHz und 16 kHz spezifiziert ist, soll für diese Arbeit generell $f_{\mathrm{A}} = 8\,\mathrm{kHz}$ angenommen werden, weil diese Abtastfrequenz für die Spracherkennung in der Praxis eine größere Verwendung findet.

Nach der ADU folgt eine Versatzkompensation sowie eine Anhebung der hohen Frequenzen, welche insgesamt eine Abflachung der spektralen Einhüllenden bewirkt. Damit soll der typische $-6\,\mathrm{dB/oct}$-Abklang des akustischen Spektrums kompensiert werden [GPAF04]. Das resultierende zeitdiskrete Signal $y(l)$, wobei $l \in \mathbb{Z}$ den Zeitindex bezeichnet, wird nun in kleine Segmente unterteilt, in denen das Signal als stationärer Zufallsprozess angesehen werden kann. Dieses geschieht durch die Multiplikation mit einer kausalen HAMMING-Analysefensterfunktion $w_{\mathrm{A}}(l)$ der endlichen Länge L_w, d. h.

$$w_{\mathrm{A}}(l) = 0 \quad \text{für} \quad l < 0 \ \wedge \ l > L_w. \tag{2.1}$$

Das Analysefenster wird dabei von einem Merkmal zum nächsten jeweils um $B \in \mathbb{N}$ Abtastwerte weitergeschoben, sodass die gefensterten Signalausschnitte

$$y_{w_{\mathrm{A}}}(m,l') := w_{\mathrm{A}}(l')y(l'+mB) \tag{2.2}$$

entstehen, wobei $m \in \mathbb{N}_0$ den Segmentindex und l' den Zeitindex für das entsprechende Segment bezeichnet. Das Analysefenster erfüllt zudem die Funktion, den bei der weiteren Berechnung des Kurzzeit-Spektrums auftretenden Leck-Effekt (engl. *leakage effect*) geeignet zu steuern [KK09].

Die gefensterten Signalsegmente werden anschließend durch die Anwendung einer diskreten FOURIER-Transformation (engl. ***D**iscrete* **F**OURIER ***T**ransform* (*DFT*)) in den Frequenzbereich transformiert, woraus das diskrete Kurzzeit-Spektrum

$$Y(m,k) = \sum_{l'=0}^{L_w-1} y_{w_\mathrm{A}}(m,l') \cdot e^{-j\frac{2\pi}{K}kl'} \tag{2.3}$$

resultiert. Dabei handelt es sich um eine in der Zeit und der Frequenz abgetastete Version der sogenannten zeitdiskreten Kurzzeit-FOURIER-Transformation (engl. ***D**iscrete-**T**ime **S**hort-**T**ime* **F**OURIER ***T**ransform* (*DTSTFT*)) [OSB99], wobei $K \in \mathbb{N}$ die Anzahl der Frequenzbins, $k \in \mathbb{N}$ den Frequenzindex und j die imaginäre Einheit bezeichnet. Da $Y(m,k)$ die Periode K bezüglich k besitzt, genügt es, nur die Indizes $k \in \{0,\ldots,K-1\}$ zu betrachten.

Anschließend wird das Kurzzeit-Leistungsspektrum gebildet und damit die Phaseninformation im Spektrum verworfen. Diese Operation wird motiviert durch perzeptuelle Studien, welche gezeigt haben, dass bei der menschlichen akustischen Wahrnehmung der Phase eine im Vergleich zur Amplitude deutlich untergeordnete Bedeutung zukommt [Gol67].

Der nächste Schritt besteht in der Berechnung der MEL-spektralen Koeffizienten $\mathscr{Y}_{m,q}$ durch eine perzeptuell motivierte Glättung des Kurzzeit-Leistungsspektrums gemäß

$$\mathscr{Y}_{m,q} := \sum_{k=K_q^{(\mathrm{u})}}^{K_q^{(\mathrm{o})}} |Y(m,k)|^2 \Lambda_q(k)\,. \tag{2.4}$$

Dabei werden überlappende Dreieckfilter $\Lambda_q(k)$, $q \in \{0,\ldots,Q-1\}$ eingesetzt, deren Zentren auf der gehörorientierten Frequenzskala, der sogenannten MEL-Frequenzskala [Kut04], äquidistant angeordnet sind. Die Abbildung zwischen tatsächlicher und wahrgenommener Frequenz verläuft bis etwa 1000 Hz näherungsweise linear und oberhalb von 1000 Hz näherungsweise logarithmisch. Die Operation, Leistungen benachbarten Frequenzbins gewichtet zusammenzufassen, ist der Eigenschaft des menschlichen Gehörs nachempfunden, die Lautstärke über Frequenzgruppen, sogenannte kritische Bänder [Gre61], gemittelt wahrzunehmen. Die Breite des q-ten Dreickfilters ergibt sich dabei jeweils aus der Differenz der oberen und unteren Grenzen $K_q^{(\mathrm{o})}$ und $K_q^{(\mathrm{u})}$.

Im Anschluss daran erfolgt eine Kompression des MEL-Spektrums durch die Anwendung des natürlichen Logarithmus zur Berechnung der **l**og-**M**EL-**s**pektralen **K**oeffizienten (LMSKs)

$$y_{m,q}^{(\mathrm{s})} := \ln\left(\mathscr{Y}_{m,q}\right). \tag{2.5}$$

Sie ist motiviert durch die Beobachtung, dass die sogenannte Lautheit, welche das Lautstärkeempfinden des Menschen widerspiegelt, sich näherungsweise logarithmisch zur tatsächlichen Schallintensität verhält. Dabei wird jedoch nicht berücksichtigt, dass das menschliche Lautstärkeempfinden frequenzabhängig ist [Kut04].

Als Folge des Überlapps der MEL-Bänder sind die LMSKs miteinander korreliert, wobei die entsprechende Kovarianzmatrix approximativ eine TOEPLITZ-Strukur aufweist. Mit Hilfe einer diskreten Kosinustransformation (engl. ***D**iscrete **C**osine **T**ransform* (*DCT*)) wird deshalb eine näherungsweise Dekorrelation durchgeführt, woraus die *MFCC*s

$$y_{m,\kappa'}^{(\mathrm{c})} := \sum_{q=0}^{Q-1} y_{m,q}^{(\mathrm{s})} \cdot \cos\left[\frac{\kappa'\pi}{Q}\left(q+\frac{1}{2}\right)\right] \tag{2.6}$$

resultieren, wobei κ' den Index und K' die Anzahl der cepstralen Komponenten angibt. Gemäß dem sogenannten Quelle-Filter-Modell lässt sich die Erzeugung eines Sprachsignals vereinfacht durch eine Faltung eines Anregungssignals mit der Impulsantwort des menschlichen Vokaltraktes beschreiben [RJ93]. Für die Spracherkennung ist jedoch nur die relativ zum Anregungssignal langsame Änderung des Vokaltraktes interessant, da diese den geformten Laut bestimmt. Deshalb werden nur *MFCC*s niederer Ordnung verwendet, was in einem kleinen Wert für K' zum Ausdruck kommt.

Bei den *MFCC*s handelt es sich um sogenannte statische Merkmale, da jeder cepstrale Koeffizient nur Auskunft über einen sehr kurzen Zeitausschnitt liefert. Die Information über einen gesprochenen Laut ist jedoch auch in der zeitlichen Änderung dieser Koeffizienten enthalten. Eine sinnvolle Ergänzung der *MFCC*s liefern die in [Fur81] eingeführten dynamischen Merkmale erster und zweiter Ordnung, die sogenannten DELTA- und DELTA-DELTA-Merkmale

$$\Delta y_{m,\kappa'}^{(\mathrm{c})} := \frac{\sum_{i=1}^{I_1} i\left(y_{m+i,\kappa'}^{(\mathrm{c})} - y_{m-i,\kappa'}^{(\mathrm{c})}\right)}{2\sum_{i=1}^{I_1} i^2} \tag{2.7}$$

$$\Delta\Delta y_{m,\kappa'}^{(\mathrm{c})} := \frac{\sum_{i=1}^{I_2} i\left(\Delta y_{m+i,\kappa'}^{(\mathrm{c})} - \Delta y_{m-i,\kappa'}^{(\mathrm{c})}\right)}{2\sum_{i=1}^{I_2} i^2}. \tag{2.8}$$

Sie stellen eine Approximation der ersten sowie zweiten Ableitung der cepstralen Merkmale nach der Zeit dar, welche durch den Segmentindex m repräsentiert wird. Die beiden Konstanten I_1 und I_2 bestimmen dabei die Größe des Zeitfensters zur Berechnung der approximativen Ableitungen. Die Hinzunahme dieser Merkmale verbessert die Erkennungsrate von Systemen zur automatischen Spracherkennung beträchtlich, was zum Teil darauf zurückzuführen ist, dass dadurch dem Erkenner für jeden Zeitausschnitt zusätzliche zeitliche Kontextinformation zur Verfügung gestellt wird.

Alle statischen und dynamischen Merkmale werden schließlich zu einem Merkmalsvektor

$$\mathbf{y}_m := \left(y_{m,0}^{(\mathrm{c})}, \ldots, y_{m,K'-1}^{(\mathrm{c})}, \Delta y_{m,0}^{(\mathrm{c})}, \ldots, \Delta y_{m,K'-1}^{(\mathrm{c})}, \Delta\Delta y_{m,0}^{(\mathrm{c})}, \ldots, \Delta\Delta y_{m,K'-1}^{(\mathrm{c})}\right)^T \tag{2.9}$$

zusammengefasst, mit Hilfe dessen die Wortsuche im Erkenner durchgeführt wird.

Abschließend sind in Tab. 2.1 die Werte der zur Merkmalsextraktion verwendeten Parameter aufgeführt.

Tabelle 2.1.: *Zur Merkmalsextraktion verwendete Parameter orientierend am ETSI-Standard [ETSb].*

Segment-länge L_w	Segment-vorschub B	Anzahl der Frequenzbins K	Anzahl der MEL-Bänder Q	Anzahl der cepstr. Koeff. K'	Einseitige Fensterlängen für dyn. Merkmale I_1	I_2
200	80	256	23	13	4	2

2.2. Decodierung

Die Decodierung ordnet einer endlichen Merkmalsvektorfolge $\mathbf{y}_{1:M} := \mathbf{y}_1, \ldots, \mathbf{y}_M$ bestehend aus M Merkmalsvektoren eine hypothetische, endliche Wortfolge $\hat{w}_{1:\hat{N}_w} := \hat{w}_1, \ldots, \hat{w}_{\hat{N}_w}$ bestehend aus $\hat{N}_w$ Wörtern zu. Dabei soll zunächst angenommen werden, dass das am Mikrophon aufgenommene Sprachsignal unverhallt und ungestört ist. Dieses wird in der hier verwendeten Notation dadurch ausgedrückt, dass die Merkmalsvektorfolge $\mathbf{y}_{1:M}$ des gewöhlich verhallten und gestörten Mikrophonsignals mit der Merkmalsvektorfolge des sauberen Sprachsignals, welche mit $\mathbf{x}_{1:M}$ bezeichnet werden soll, gleichgesetzt wird. Es gilt daher $\mathbf{y}_{1:M} = \mathbf{x}_{1:M}$.

Die Bestimmung der Wortfolge $\hat{w}_{1:\hat{N}_w}$ erfolgt gemäß der BAYES'schen Entscheidungsregel

$$\hat{w}_{1:\hat{N}_w} = \underset{N_w, w_{1:N_w}}{\operatorname{argmax}} P_{\breve{w}_{1:N_w}|\breve{\mathbf{x}}_{1:M}}\left(w_{1:N_w}|\mathbf{x}_{1:M}\right), \tag{2.10}$$

wobei $\breve{w}_{1:N_w}$ und $\breve{\mathbf{x}}_{1:M}$ die der Wortfolge $w_{1:N_w}$ und der Merkmalsvektorfolge $\mathbf{x}_{1:M}$ zugrunde liegenden Zufallsprozesse bezeichnen und $P_{\breve{w}_{1:N_w}|\breve{\mathbf{x}}_{1:M}}$ die auf $\breve{\mathbf{x}}_{1:M}$ bedingte Wahrscheinlichkeitsmassefunktion von $\breve{w}_{1:N_w}$ darstellt. Im Sinne einer verbesserten Lesbarkeit werden im Folgenden die Subskripte von Wahrscheinlichkeitsmasse- und Verteilungsdichtefunktionen überall dort weggelassen, wo die jeweilige Zufallsvariable oder der jeweilige Zufallsprozess offensichtlich aus dem Argument der entsprechenden Funktion erkennbar wird. Damit lässt sich (2.10) verkürzt auch gemäß

$$\hat{w}_{1:\hat{N}_w} = \underset{N_w, w_{1:N_w}}{\operatorname{argmax}} P\left(w_{1:N_w}|\mathbf{x}_{1:M}\right) \tag{2.11}$$

darstellen. Durch die Anwendung der BAYES'schen Regel für bedingte Wahrscheinlichkeiten lässt sich (2.11) wie folgt formulieren:

$$\hat{w}_{1:\hat{N}_w} = \underset{N_w, w_{1:N_w}}{\operatorname{argmax}} \frac{p\left(\mathbf{x}_{1:M}|w_{1:N_w}\right) P\left(w_{1:N_w}\right)}{p\left(\mathbf{x}_{1:M}\right)} \tag{2.12}$$

$$= \underset{N_w, w_{1:N_w}}{\operatorname{argmax}} p\left(\mathbf{x}_{1:M}|w_{1:N_w}\right) P\left(w_{1:N_w}\right). \tag{2.13}$$

wobei für die Umformung (2.12) die BAYES'sche Regel für bedingte Wahrscheinlichkeiten verwendet wurde und in (2.13) schließlich ausgenutzt wurde, dass der Term $p\left(\mathbf{x}_{1:M}\right)$ für die Maximierung irrelevant ist.

Man erkennt, dass für die Lösung der Decodieraufgabe die Verteilungsdichtefunktionen $p_{\breve{\mathbf{x}}_{1:M}|\breve{w}_{1:N_w}}$ und die Wahrscheinlichkeitsmassefunktionen $P_{\breve{w}_{1:N_w}}$ benötigt werden, welche jeweils parametrisch durch das *akustische Modell* und das *Sprachmodell* beschrieben werden. In der Praxis wird das Sprachmodell oft mit einer empirisch bestimmten Konstanten $\alpha^{(\mathrm{SM})}$

skaliert, um dem Sprachmodell gegenüber dem akustischen Modell mehr Gewicht zu verleihen. Dadurch resultiert eine im Vergleich zu (2.13) etwas veränderte Decodiervorschrift

$$\hat{w}_{1:\hat{N}_w} = \underset{N_w, w_{1:N_w}}{\operatorname{argmax}} \, p\left(\mathbf{x}_{1:M} | w_{1:N_w}\right) P^{\alpha^{(\mathrm{SM})}}\left(w_{1:N_w}\right). \tag{2.14}$$

Akustisches Modell

Das akustische Modell nimmt an, dass der beobachteten Merkmalsvektorfolge $\mathbf{x}_{1:M}$ eine von der entsprechenden Wortfolge $w_{1:N_w}$ abhängige, jedoch verborgene Zustandssequenz $\gamma_{1:M}$ zugrunde liegt. Diese wird wiederum als Realisierung eines Zufallsprozesses $\breve{\gamma}_{1:M}$ betrachtet, um damit Variationen in der Aussprache der Wortfolge Rechnung zu tragen. Mit dem Gesetz der totalen Wahrscheinlichkeit kann die Verteilungsdichtefunktion $p\left(\mathbf{x}_{1:M} | w_{1:N_w}\right)$ dann gemäß

$$p\left(\mathbf{x}_{1:M} | w_{1:N_w}\right) = \sum_{\{\gamma_{1:M}\}} p\left(\mathbf{x}_{1:M} | \gamma_{1:M}, w_{1:N_w}\right) P\left(\gamma_{1:M} | w_{1:N_w}\right) \tag{2.15}$$

dargestellt werden, wobei die Summation über alle möglichen Zustandssequenzen $\gamma_{1:M}$ zu bilden ist. Im Sinne einer handhabbaren Auswertung der Verteilungsdichtefunktion (2.15) werden anschließend zwei einschneidende Annahmen gemacht.

Zum einen wird der Zufallsprozess $\breve{\gamma}_{1:M}$ als eine diskrete, endliche MARKOV-Kette erster Ordnung [RJ93] modelliert, woraus auch die Bezeichnung HMM für das akustische Modell resultiert. Gemäß dieser Modellierung hängt die Wahrscheinlichkeit, dass $\breve{\gamma}_m$ einen bestimmten Wert γ_m annimmt, nur vom Wert γ_{m-1} der Zufallsvariable $\breve{\gamma}_{m-1}$ ab.

Zum anderen wird davon ausgegangen, dass ein Merkmalsvektor $\mathbf{x}_m$ mit dem Segmentindex m nur vom Zustand γ_m, jedoch insbesondere nicht von vorhergehenden oder nachfolgenden Merkmalsvektoren, abhängt. Diese im Englischen unter dem Begriff *conditional independence assumption* weit verbreitete Annahme modelliert sämtliche Abhängigkeiten zwischen den Merkmalsvektoren nur über den Zustandsprozess $\breve{\gamma}_{1:M}$. In ihr besteht auch der größte Schwachpunkt der Modellierung, da mit Hilfe der MARKOV-Kette nur ein relativ begrenzter zeitlicher Kontext erfasst wird. Mit der Hinzunahme von in Kap. 2.1 eingeführten dynamischen Merkmalen wird versucht, diesem Problem teilweise entgegen zu wirken.

Unter den beiden genannten Voraussetzungen lässt sich (2.15) durch

$$p\left(\mathbf{x}_{1:M} | w_{1:N_w}\right) \approx \sum_{\{\gamma_{1:M}\}} \prod_{m=1}^{M} p\left(\mathbf{x}_m | \gamma_m, w_{1:N_w}\right) P\left(\gamma_m | \gamma_{m-1}, w_{1:N_w}\right) \tag{2.16}$$

approximieren. Dabei beschreiben die Wahrscheinlichkeiten $P\left(\gamma_m | \gamma_{m-1}, w_{1:N_w}\right)$, $1 \leq m \leq M$, die auf die Wortfolge $w_{1:N_w}$ bedingten Zustandsübergänge. Entsprechende Wahrscheinlichkeiten basieren auf dem Konzept, dass zunächst abhängig von der Größe des Lexikons einzelne HMMs für Wörter oder Wortuntereinheiten aufgestellt und anschließend sinnvoll konkateniert werden. Als Wortuntereinheiten werden meist die sogenannten Triphone verwendet. Darunter versteht man kontextabhängige Phoneme, welche von ihrem Vorgänger- und Nachfolgephonem bestimmt werden. Die zustandsbedingten Verteilungsdichtefunktionen $p\left(\mathbf{x}_m | \gamma_m, w_{1:N_w}\right)$, $1 \leq m \leq M$, die auch als Emissionsverteilungsdichtefunktionen bezeichnet werden, werden in der Regel durch GAUSS-Mischungsmodelle (engl. **G**AUSSIAN ***M**ixture **M**odel*s (*GMM*s)) beschrieben.

Die Parameter des akustischen Modells werden mit Hilfe von Trainingsdaten, welche aus gesprochenen Äußerungen in Form von akustischen Signalen sowie deren Transkription bestehen, mit dem ***E**xpectation **M**aximization* (*EM*)-Algorithmus [RJ93] geschätzt. Man spricht dabei auch von überwachtem Training, da die Transkription bekannt ist.

Sprachmodell

Das Sprachmodell ist typischerweise ein $N^{(\mathrm{SM})}$-Gram, was bedeutet, dass die Auftrittswahrscheinlichkeit eines Wortes nur von den $N^{(\mathrm{SM})}-1$ vorhergehenden Wörtern abhängt. Die Wahrscheinlichkeit für das Auftreten einer bestimmten Wortfolge $w_{1:N_w}$ lässt sich damit durch

$$P\left(w_{1:N_w}\right) \approx \prod_{\nu=1}^{N_w} P\left(w_\nu | w_{\nu-N^{(\mathrm{SM})}:\nu-1}\right) \tag{2.17}$$

annähern. Die zur Auswertung des rechten Terms benötigten bedingten Wortwahrscheinlichkeiten $P\left(w_\nu | w_{\nu-N^{(\mathrm{SM})}:\nu-1}\right)$ werden in der Trainingsphase unter Verwendung von reinen Textdatenbanken geschätzt, indem jeweils die relative Häufigkeit des Auftretens des Wortes w_ν nach der Wortfolge $w_{\nu-N^{(\mathrm{SM})}:\nu-1}$ bestimmt wird.

Als Ergebnis der durch das akustische und das Sprachmodell eingeführten Approximationen (2.16) und (2.17) erfolgt die Decodierung nach der vereinfachten Regel

$$\hat{w}_{1:\hat{N}_w} = \underset{N_w, w_{1:N_w}}{\operatorname{argmax}} \sum_{\{\gamma_{1:M}\}} \prod_{m=1}^{M} p\left(\mathbf{x}_m | \gamma_m, w_{1:N_w}\right) P\left(\gamma_m | \gamma_{m-1}, w_{1:N_w}\right) \prod_{\nu=1}^{N_w} P\left(w_\nu | w_{\nu-N^{(\mathrm{SM})}:\nu-1}\right). \tag{2.18}$$

Mit einer weiteren Vereinfachung, bei der die Summation durch die Maximumbildung über alle möglichen Zustandssequenzen ersetzt wird, lässt sich die Maximierungsaufgabe sehr effizient mit dem VITERBI-Algorithmus [RJ93] lösen. Eine zusätzliche Operation, bei der sehr viele Rechenoperationen eingespart werden können, ist das frühzeitige Verwerfen (engl. *pruning*) bestimmter Wort- bzw. Zustandskombinationen, falls diese zu unwahrscheinlich werden. Damit kann eine erhebliche Einschränkung des Suchraums erreicht werden, wobei zu berücksichtigen ist, dass das Ergebnis dann im Allgemeinen nur suboptimal ist.

2.3. Spracherkennung in halligen Umgebungen

Bedingt durch den vergrößerten Abstand des Sprechers zum Mikrophon bei der Verwendung einer Freisprecheinrichtung muss das Sprachsignal in Form von Schallwellen einen größeren direkten Weg von der Quelle zur Senke zurücklegen, so dass es einerseits eine Dämpfung durch die Energieabsorption durch das Medium erfährt. Andererseits gelangt das Signal nicht nur über den direkten Pfad vom Sprecher zum Mikrophon, sondern auch über Umwege, welche sich aus Reflexionen der Schallwellen an Oberflächen von Wänden oder Gegenständen ergeben. Die daraus resultierenden verzögerten und gedämpften Versionen des Sprachsignals überlagern das eigentliche Sprachsignal und werden als Nachhall wahrgenommen [Kut04, Kap. 4]. Zusätzlich beinhaltet das Mikrophonsignal in der Regel Hintergrundstörungen, welche zum Teil auch aus der Sprache konkurrierender Sprecher bestehen können.

Das Mikrophonsignal lässt sich vereinfacht gemäß

$$y(l) = s(l) + n(l) \tag{2.19}$$

darstellen, wobei $n(l)$ das Störsignal und $s(l)$ das verhallte Sprachsignal bezeichnet. Das letztere kann vereinfacht durch eine Faltung des sauberen Sprachsignals $x(l)$ mit einer sogenannten **R**aum**i**mpuls**a**ntwort (RIA) $h(l)$ gemäß

$$s(l) = (x * h)(l) \tag{2.20}$$

beschrieben werden, wobei die RIA das Übertragungsverhalten der Umgebung vom Sprecher zum Mikrophon charakterisiert. Die Vereinfachung bei dieser Darstellung besteht in der Annahme einer zeitinvarianten RIA, welche in der Regel nicht gerechtfertigt ist wie aus den folgenden Ausführungen deutlich wird.

Eine beispielhafte RIA, welche in einem großen Büro gemessen wurde, ist in Abb. 2.3 dargestellt. Sie lässt sich typischerweise grob in drei Bereiche einteilen, die auf einer geo-

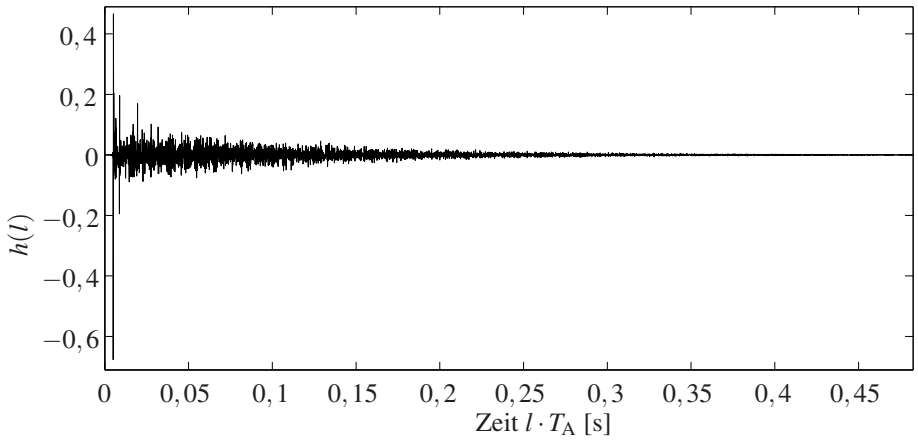


Abbildung 2.3.: *Beispielhafte Raumimpulsantwort gemessen in einem großen Büro ($T_{60} \approx 0{,}75\,\mathrm{s}$, $\mathrm{DRR} \approx 0\,\mathrm{dB}$).*

metrischen Interpretation beruhen [Kut04, Kap. 4]. Der erste ist durch den direkten Anteil gegeben, der sich in dem ersten verzögerten Impuls mit einer verhältnismäßig großen Amplitude äußert. Der zweite Bereich besteht aus einigen sporadisch auftretenden und stärker gedämpften Impulsen, welche von signifikanten frühen Reflexionen herrühren. Die unterschiedlichen Vorzeichen der einzelnen Impulse entstehen durch Phasensprünge, welche bei Reflexionen stattfinden. Die temporale Dichte der Impulse vergrößert sich quadratisch mit der Zeit, so dass sich diese nach einiger Zeit zwangsläufig überlagern und die Anzahl der gleichzeitig überlagerten Impulse im Mittel weiter zunimmt. So ist der dritte Bereich, der ab etwa 50 ms nach dem Hauptimpuls beginnt, durch scheinbar zufällig auftretende, aufeinanderfolgende Impulse gekennzeichnet, die näherungsweise als Stichproben von unabhängigen GAUSS-verteilten Zufallsvariablen interpretiert werden können. Dabei nimmt die Energie der späten Reflexionen approximativ exponentiell mit der Zeit ab, was grob, aber anschaulich, dadurch erklärt werden kann, dass bei jeder stattfindenden Reflexion ein gewisser Anteil der Energie der Schallwelle absorbiert wird. Das Abklingverhalten der Energie lässt sich mit

Hilfe der sogenannten Energieabfallkurve (engl. ***E**nergy **D**ecay **C**urve* (*EDC*)) beschreiben, die durch eine normierte Rückwärtsintegration der quadratischen Raumimpulsantwort wie folgt berechnet werden kann:

$$\mathrm{EDC}_h(l) := \frac{\sum_{p'=l}^{\infty} h^2(p')}{\sum_{p'=0}^{\infty} h^2(p')}. \tag{2.21}$$

Abbildung 2.4 zeigt die zur RIA in Abb. 2.3 gehörige *EDC* in einer logarithmischen Darstellung. Erwartungsgemäß lässt sich eine affine Zeitabhängigkeit für den Bereich der späten

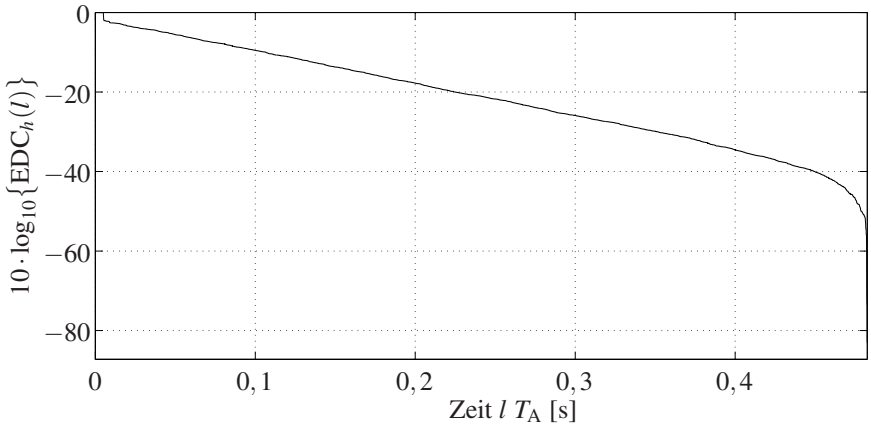


Abbildung 2.4.: *Energieabfallkurve (in einer logarithmischen Darstellung) zur Raumimpulsantwort in Abb. 2.3.*

Reflexionen beobachten.

Eine wesentliche Größe zur Charakterisierung von Räumen bzw. RIAs ist die sogenannte Nachhallzeit T_{60}. Sie ist definiert als diejenige Zeit, welche benötigt wird, damit die Energie des (eigentlich späten) Nachhalls um 60 dB gegenüber dem initialen Wert abklingt [Kut04, Kap. 5]. Sie lässt sich gemäß [Sch65] aus der Steigung der logarithmierten *EDC* bestimmen. Bemerkenswert ist weiterhin die Tatsache, dass die Energie der frühen Reflexionen ebenfalls exponentiell abklingt, jedoch manchmal mit einer anderen Abklingkonstanten. Dies führt dazu, dass der Verlauf der logarithmierten *EDC* nicht mehr affin, sondern nur noch stückweise affin ist [Sch65].

Weiterhin ist zu beachten, dass sich das Abklingverhalten der Energie im Allgemeinen frequenzabhängig ist. Diese Eigenschaft ist bedingt durch die Tatsache, dass Materialen die Energie von Schallwellen unterschiedlicher Frequenzen unterschiedlich stark absorbieren. In der Regel werden hochfrequente Anteile von Materialen stärker gedämpft als tieffrequente, so dass die Energie der tieffrequenten Anteile langsamer abklingt. Dieses Phänomen wird bei der Bestimmung der Nachhallzeit aus der *EDC* nach dem zuvor beschriebenen Prinzip nicht berücksichtigt.

Während die Nachhallzeit sehr grob die Eigenschaft eines Raumes beschreibt, liefert sie keine Auskunft über die Konfiguration des Sprechers und des Mikrophons innerhalb des Raumes. Eine Möglichkeit einer qualitativen Charakterisierung des Abstandes beider bietet das Verhältnis zwischen der Energie des direkten Schallanteils und der Energie des Nachhalls

einschließlich der frühen Reflexionen (engl. ***D**irect-to-**R**everberant **R**atio* (*DRR*)) , welches durch

$$\mathrm{DRR} := 10\log_{10}\left(\frac{\sum_{l=0}^{l_\mathrm{D}} h^2(l)}{\sum_{l=l_\mathrm{D}+1}^{\infty} h^2(l)}\right)\,[\mathrm{dB}] \tag{2.22}$$

definiert ist [Hab07]. Dabei wird angenommen, dass der Zeitindex l_D jenem Zeitpunkt entspricht, an dem der Hauptimpuls auftritt. Bei gemessenen RIAs ist die präzise Bestimmung des Hauptimpulses meist nicht möglich, so dass der Wert von $l_\mathrm{D} \cdot T_\mathrm{A}$ oft so gewählt wird, dass er 8 –16 ms größer als die Ankunftszeit des direkten Schalls ist. Dabei wird in dieser Arbeit stets von 10 ms ausgegangen, falls Werte des *DRR* angegeben werden.

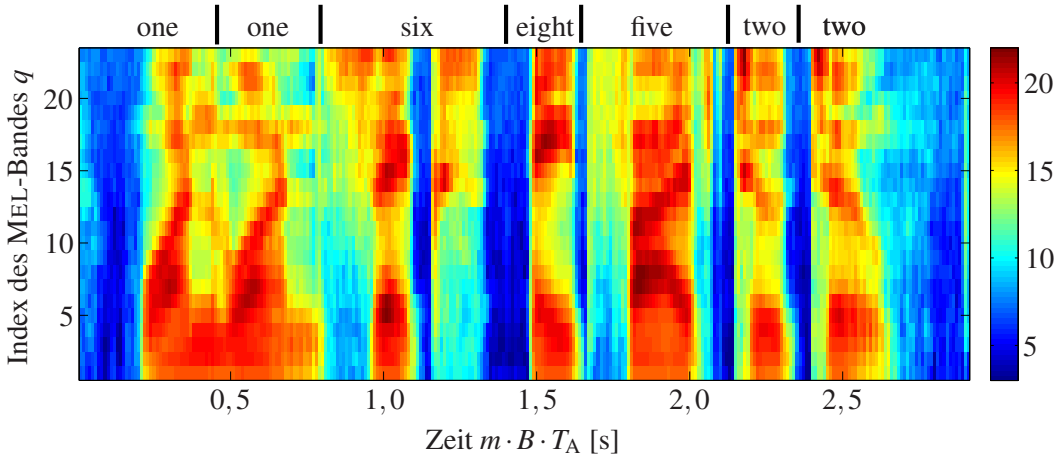
Neben dem *DRR* gibt es zahlreiche weitere Maße wie z. B. das Klarheitsmaß C_{50} bzw. C_{80}, welche zur Beschreibung der Auswirkungen der RIA auf die Verständlichkeit von Sprache bzw. die Durchsichtigkeit von Musik für den Menschen verwendet werden können. Eine ausführliche Übersicht über in der Literatur auftretende subjektive und objektive Maße zur Bestimmung des Einflusses des Nachhalls findet sich beispielsweise in [Ric09, Kap. 4.4]. Auf eine detaillierte Darstellung dieser Maße wird hier verzichtet, da die Auswirkungen des Nachhalls auf die Spracherkennung im Vordergrund stehen und mit den beiden Größen T_{60} und DRR bereits eine in diesem Zusammenhang vernünftige und in der Literatur übliche qualitative Beschreibung der RIA gegeben ist.

Typischerweise ist die RIA in hohem Maße zeitvariant, was unter anderem auf Bewegungen des Sprechers sowie bereits geringe Änderungen der Temperatur und Feuchtigkeit innerhalb des Raumes zurückgeführt werden kann. Diese Änderungen betreffen jedoch in der Regel den Direktanteil, die frühen Reflexionen sowie im Allgemeinen die feine Struktur der RIA. Hingegen wird die grobe Charakteristik, mit der hier die Einhüllende des späten Nachhalls sowie die Nachhallzeit T_{60} gemeint ist, dadurch kaum beeinflusst.

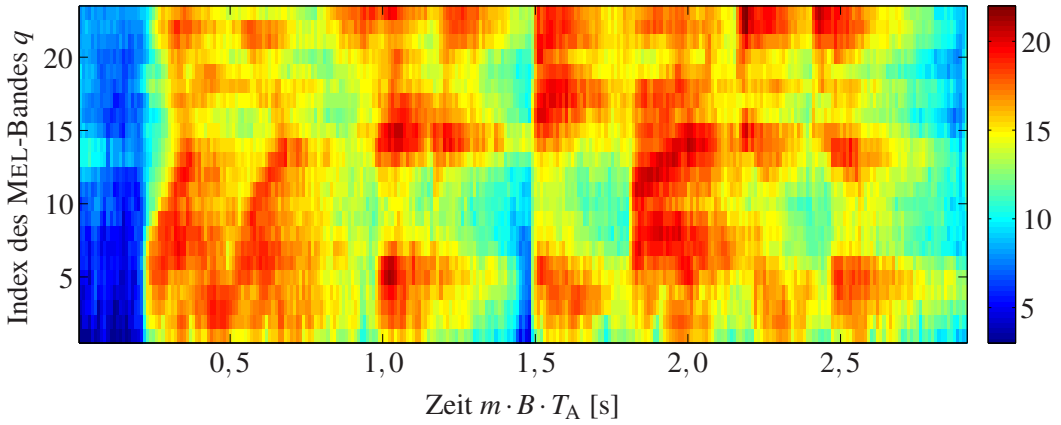
Die Auswirkungen des Nachhalls auf das Sprachsignal $x(l)$ sind zweierlei. Während die frühen Reflexionen zu einer sogenannten Färbung (engl. *coloration*) des Kurzzeit-Spektrums führen [Kut04], bewirkt der späte Nachhall im Wesentlichen eine zeitliche Dispersion des Sprachsignals, die sich entsprechend in der Trajektorie der log-MEL-spektralen Merkmale wiederfinden lässt. Dieser Effekt wird beispielhaft in Abb. 2.5, die die Trajektorien der log-MEL-spektralen Merkmale einer sauberen und verhallten Version eines Sprachsignals zeigt, veranschaulicht. Die zugehörige Sprachäußerung wurde der AURORA5-Datenbank entnommen und entspricht der in amerikanischem Englisch ausgesprochenen Ziffernkette "*one*, *one*, *six*, *eight*, *five*, *two*, *two*". Zur Verhallung wurde die konstante RIA aus Abb. 2.3 verwendet. Bei dem Vergleich der Trajektorien fällt zum Beispiel auf, dass der Glottalschlag (engl. *glottal stop*) bei der Aussprache der Ziffer "*six*" bei etwa 1,2 s, der in Abb. 2.5a sehr deutlich zu erkennen ist, in Abb. 2.5b vollkommen durch den Nachhall verdeckt ist.

Die durch den Nachhall verursachte zeitliche Dispersion innerhalb der Trajektorie der log-MEL-spektralen Merkmale des Sprachsignals führt offensichtlich zu einer Änderung ihrer statistischen Eigenschaften, damit zu einer Diskrepanz zwischen den Trainings- und Testbedingungen und letztendlich zu einer Erhöhung der Wortfehlerrate des Spracherkenners. Zusätzlich werden die statistischen Eigenschaften der Merkmale durch Hintergrundstörungen beeinflusst, was jedoch in einer grundsätzlich unterschiedlichen Art geschieht. Denn während für Hintergrundstörungen oft die Annahme gerechtfertigt ist, dass sie keine Korrelation

zum Sprachsignal aufweisen, besteht zwischen dem Nachhall und dem sauberen Sprachsignal eine starke Korrelation. Aufgrund dessen bewirkt der Nachhall eine stärkere Verletzung der Annahme über die gegenseitige bedingte Unabhängigkeit von zeitlich aufeinanderfolgenden Merkmalsvektoren (siehe Kap. 2.2). Vergleicht man folglich die Leistungsfähigkeit von HMM-basierten Spracherkennern unter Trainings- und Testbedingungen, welche auf der Verwendung von sauberen Sprachsignalen einerseits und verhallten Sprachsignalen andererseits basieren, so wird diese in der Regel im zweiten Fall schlechter ausfallen.



(a) Trajektorie der log-MEL-spektralen Merkmale $x^{(\mathrm{s})}_{m,q}$ des sauberen Sprachsignals samt der entsprechenden Transkription



(b) Trajektorie der log-MEL-spektralen Merkmale $s^{(\mathrm{s})}_{m,q}$ des verhallten Sprachsignals

Abbildung 2.5.: *Trajektorien der log-*MEL*-spektralen Merkmale einer sauberen und verhallten Version eines beispielhaften Sprachsignals zugehörig zu der Ziffernkettenäußerung "one, one, six, eight, five, two, two". Zur künstlichen Verhallung wurde die RIA aus Abb. 2.3 verwendet.*

3. Stand der Forschung

Die in der Literatur bisher existenten Verfahren zur hall- und störrobusten Spracherkennung lassen sich grob in drei Kategorien unterteilen. Diese unterscheiden sich dadurch, dass sie jeweils an einer anderen Stelle innerhalb eines Spracherkennungssystems zum Einsatz kommen. Während die signalbasierten Verfahren bestrebt sind, das Sprachsignal bereits vor der Merkmalsextraktion zu enthallen und zu entstören, besteht das Ziel der merkmalsbasierten Verfahren in einer robusten Extraktion der akustischen Merkmale. Dazu gehört auch eine sinnvolle Manipulation bereits extrahierter Merkmale im Hinblick auf deren Enthallung und Entstörung. Die dritte Kategorie besteht aus den Methoden zur Anpassung des akustischen Modells oder des Decoders an den Nachhall oder die Hintergrundstörung. Im Folgenden werden die drei Kategorien detailliert vorgestellt. Dabei beschränkt sich der Überblick fast ausschließlich auf die Verfahren, die im Zusammenhang mit der Robustheit gegenüber dem Nachhall stehen, da dieses Thema den Schwerpunkt der Arbeit darstellt.

3.1. Verfahren zur Enthallung des akustischen Signals

Das Hauptaugenmerk der signalbasierten Verfahren liegt auf der Rekonstruktion des sauberen Sprachsignals aus dem verhallten und gestörten Sprachsignal. Im Hinblick auf eine hall- und störrobuste Spracherkennung lässt sich das mit derartigen Methoden rekonstruierte Signal anschließend einer Merkmalsextraktion unterziehen. Dabei sei betont, dass die Spracherkennung nicht die einzige Anwendung für die signalbasierte Enthallung und -entstörung darstellt. So steht zum Beispiel für viele solcher Verfahren die Verbesserung der Sprachverständlichkeit für den Menschen im Vordergrund. Da die signalbasierten Verfahren nicht den Fokus dieser Dissertation bilden, wird an dieser Stelle nur ein sehr kurzer Überblick über diese gegeben, der keinen Anspruch auf Vollständigkeit erhebt. Für weitere Details sei der Leser auf die ausführlicheren Übersichten in [Hab07, Kap.3] und [HBC08] verwiesen.

Grundsätzlich lassen sich die signalbasierten Verfahren danach unterscheiden, ob ihr Ziel darin besteht, den Nachhall vollständig zu entfernen oder aber nur zu unterdrücken. Weiterhin unterscheidet man zwischen ein- oder mehrkanaligen Methoden sowie dem Grad des verwendeten A-priori-Wissens über das Sprachsignal oder die Umgebung, in der das Sprachsignal aufgenommen wird [Hab07, Kap.3].

3.1.1. Verfahren zur Entfernung des Nachhalls

Einige der Verfahren zur Entfernung des Nachhalls verfolgen die Idee, einen Entzerrer auf das verhallte Sprachsignal anzuwenden, welcher den Effekt der Faltung mit der RIA rückgängig macht. Da die RIA in gewöhnlichen Anwendungen unbekannt ist, muss diese zu-

nächst aus dem verhallten Sprachsignal geschätzt werden. Ein wesentliches Problem dieses Ansatzes ist die Tatsache, dass die exakte Inversion der RIA im Allgemeinen nicht möglich ist. Denn dazu ist erforderlich, dass die zeitdiskrete RIA minimalphasig ist, was bedeutet, dass die Nullstellen ihrer z-Transformierten innerhalb des Einheitskreises in der komplexen Ebene liegen. Dieses trifft jedoch für typische Räume wie Büros und Wohnzimmer gewöhnlich nicht zu [NA79], so dass theoretisch nur eine approximative Inversion der RIA gelingen kann. Außerdem liegen die Nullstellen der z-Transformierten der RIA oft nahe dem Einheitskreis, so dass Stabilitätsprobleme bedingt durch die Approximationen bei der numerischen Umsetzung der Inversion auftreten können. Insbesondere weist das inverse Filter eine hohe Sensibilität gegenüber kleinen Änderungen der RIA auf [Mou85, RWK00, TW02], welche sowohl durch die Zeitvarianz der RIA bedingt durch beispielsweise geringe Bewegungen des Sprechers als auch durch Fehlschätzungen der RIA verursacht werden können.

Unter der Voraussetzung, dass mehrere Mikrophone für die Aufnahme der akustischen Signale zur Verfügung stehen, konnte in [MK88] gezeigt werden, dass trotz der fehlenden Minimalphasigkeit der RIAs deren exakte Inversion unter relativ milden Bedingungen möglich ist. Diese Aussage bildet den Kern des sogenannten ***M**ultiple Input/Output **IN**verse **T**heorem* (*MINT*). Eine notwendige Bedingung besteht zum Beispiel darin, dass die z-Transformierten der zeitdiskreten RIAs vom Sprecher zu sämtlichen Mikrophonen keine gemeinsamen Nullstellen besitzen dürfen [MK88].

Für die Schätzung der RIA lassen sich eigenraumbasierte Verfahren nutzen, welche mehrere Mikrophone erfordern [GN95, GM03, Gan08, Gan10]. Dabei wird die RIA aus dem Nullraum einer aus den Abtastwerten aller Mikrophonsignale konstruierten Korrelationsmatrix extrahiert. Die Schätzung der RIA wird dabei insbesondere durch ihre Zeitvarianz sowie die Präsenz von Hintergrundstörungen erschwert. Um den Einfluss der Fehler in der geschätzten RIA auf die Bestimmung des inversen Filters zu reduzieren, wurde in [HDM06] eine Regularisierung vorgeschlagen. Obwohl dadurch die Sensitivität gegenüber Schätzfehlern reduziert wird, ist die erzielte Entzerrung nur suboptimal.

Eine weitere Möglichkeit zur Entfernung des Nachhalls unter der Voraussetzung der Präsenz mehrerer Mikrophonsignale besteht in der direkten Entfaltung des verhallten Sprachsignals, welche ohne die explizite Schätzung der RIA auskommt [TS05, DHM07, NYK$^+$08]. Es basiert auf der Anwendung von linearer Prädiktion (engl. ***L**inear **P**rediction* (*LP*)) zur blinden Entzerrung. Ein unerwünschter Effekt des Entzerrers ist die gleichzeitige Entzerrung bezüglich der zeitvarianten Übertragungsfunktion des Vokaltraktes, welche für die Lautformung gemäß dem Quelle-Filter-Modell der Sprachsignalerzeugung [RJ93, Kap. 3.3] wesentlich ist. Zur Kompensation dieses Effektes muss die Übertragungsfunktion des Vokaltraktes mit geschätzt werden, so dass ein entsprechendes inverses Filter bestimmt werden kann. Um die im Vergleich zur RIA relativ kurze Impulsantwort des Vokaltraktes bei der linearen Prädiktion außer Acht zu lassen, können alternativ Verfahren wie die mehrstufige lineare Prädiktion (engl. ***M**ulti-**S**tep **L**inear **P**rediction* (*MSLP*)) [GD97, KDNM09] verwendet werden. Dabei wird ein Abtastwert nicht durch seine direkten Vorgänger vorhergesagt, sondern durch einige weiter zurückliegende, aufeinander folgende Abtastwerte. Dadurch werden also nur durch den späten Nachhall eingeführten Korrelationen im Sprachsignal berücksichtigt.

3.1.2. Verfahren zur Unterdrückung des Nachhalls

In Abgrenzung zu den eben erläuterten Verfahren zur vollständigen Entfernung des Nachhalls existieren in der Literatur zahlreiche Ansätze zur Unterdrückung des Nachhalls. Dazu gehört beispielsweise die Unterdrückung des späten Nachhalls mit Hilfe von spektraler Subtraktion [LBD01, Hab04, PS06], wobei die Verbesserung der Sprachverständlichkeit das primäre Ziel darstellt. Dabei wird davon ausgegangen, dass die späten Reflexionen unkorreliert zum direkten Anteil samt den frühen Reflexionen sind. Mit dieser Argumentation kann der späte Nachhall als zum gewünschten Sprachsignal unkorrelierte, additive Störung aufgefasst werden, so dass Methoden zur Störunterdrückung basierend auf der spektralen Subtraktion angewendet werden können. Die Herausforderung bei solchen Methoden stellt die akkurate Schätzung des Kurzzeit-Leistungsdichtespektrums des späten Nachhalls dar, für welche ein statistisches Modell der RIA herangezogen wird, das die Nachhallzeit T_{60} als einzigen Parameter besitzt. Es ist bei diesem Verfahren ebenfalls möglich, frequenzabhängige Nachhallzeiten zur genaueren Modellierung des Absorptionseigenschaften der Oberflächen von umgebenden Wänden und Objekten zu berücksichtigen [Hab04].

Alternative Ansätze zur Unterdrückung des Nachhalls basieren auf der Anwendung der Modulationstheorie auf Sprachsignale. Beispielsweise wird in [HNKT00, UFSA03] ein sauberes Sprachsignal als Produkt eines weißen GAUSS'schen Zufallsprozesses und einer Einhüllenden modelliert. Eine ähnliche Modellierung wird für die RIA vorgenommen, wobei eine exponentiell abklingenden Einhüllende zugrunde gelegt wird. Aufbauend darauf lässt sich die Einhüllende eines verhallten Sprachsignals durch die Faltung der Einhüllenden des sauberen Sprachsignals und der der RIA ausdrücken. Diese Operation führt zu einer Verringerung der Modulationstiefe, deren Ausmaß mit Hilfe einer im Englischen als ***M**odulation **T**ransfer **F**unction* (*MTF*) bezeichneten Übertragungsfunktion charakterisiert werden kann [HS85]. Die Verringerung der Modulationstiefe insbesondere im Bereich der Modulationsfrequenzen zwischen etwa 0,5 Hz und 20 Hz geht dabei mit der Verschlechterung der Sprachverständlichkeit einher [HS85]. Zur Rekonstruktion der Einhüllenden des sauberen Sprachsignals muss folglich eine inverse Filterung der Einhüllenden des verhallten Sprachsignals durchgeführt werden. Es existieren dabei auch Methoden, welche das Sprachsignal innerhalb einzelner kritischer Bänder als amplitudenmoduliertes Signal auffassen und dementsprechend eine Verbesserung von bandspezifischen Einhüllenden vornehmen [LS82, MH83].

In Abgrenzung dazu gibt es Verfahren, die die Enthallung durch eine Verbesserung des Residuums, welches bei der Vorhersage eines Sprachsignals durch lineare Prädiktion entsteht, realisieren. Dabei wird das sogenannte *LP*-Residuum, was aus einer *LPC*-Analyse von kurzen Segmenten des Sprachsignals resultiert, zunächst geeignet modifiziert, um daraus anschließend das verbesserte Sprachsignal zu resynthetisieren. Grundsätzlich repräsentiert das *LP*-Residuum gemäß dem Quelle-Filter-Modell der Sprachsignalerzeugung [RJ93, Kap. 3.3] das Anregungssignal, welches durch den Vokaltrakt geformt wird. Daher werden innerhalb von Bereichen, die stimmhaften Lauten entsprechen, Glottalschläge im *LP*-Residuum als abschnittsweise periodisch auftretende Spitzen sichtbar. Durch den Einfluss des Nachhalls werden diese über die Zeit verschmiert. Unter der wesentlichen Annahme, dass die *LPC*-Koeffizienten durch den Nachhall nicht verändert werden, geschieht beispielsweise in [YM00] die Modifikation des *LP*-Residuums derart, dass versucht wird, die dem direkten Anteil entsprechenden Spitzen zu verstärken. Ein ähnlich motivierter Ansatz wird in [GMF01] verfolgt. Das Ausmaß der Verschmierung der Spitzen im *LP*-Residuum hängt

direkt mit der Intensität des Nachhalls zusammen. Dabei verringert sich mit zunehmender Intensität des Nachhalls die Kurtosis des *LP*-Residuums. Ausgehend davon wird in [GMF01] versucht, eine adaptive Filterung des *LP*-Residuums derart durchzuführen, dass die Kurtosis des gefilterten Signals maximiert wird.

Die Annahme, dass die *LPC*-Koeffizienten durch den Nachhall nicht verändert werden, trifft jedoch im Allgemeinen nicht zu. Eine solche Aussage ist nämlich nur gültig für den Erwartungswert der *LPC*-Koeffizienten bezüglich aller räumlichen Positionen des Sprechers und des Mikrophons, nicht jedoch für eine feste Anordnung beider. Dieses konnte mit der Verwendung der statistischen Raumakustik [Kut00] in [GNW03] gezeigt werden. Als Folge dessen wird zur genaueren Bestimmung der *LPC*-Koeffizienten in [GNW03, GRTN10] vorgeschlagen, mehrere Mikrophone zur Aufnahme des Sprachsignals zu verwenden, um anschließend die auf jedem einzelnen Signal bestimmten *LPC*-Koeffizienten zu mitteln.

Sehr ähnlich dazu sind Verfahren, welche A-priori-Information über die Sprache in Form ihrer harmonischen Struktur ausnutzen [NM03, KNM05, NJKM05, NMK05, NKM07]. Dabei werden Schätzungen der Stimmbandgrundfrequenz sowie der harmonischen Struktur des Sprachsignals dazu verwendet, den direkten Anteil des verhallten Sprachsignals zu rekonstruieren. Auch diese Methoden nehmen an, dass sich die Stimmbandgrundfrequenz durch den Einfluss des Nachhalls nicht verändert und sich deshalb robust aus einem verhallten Sprachsignal schätzen lässt.

Eine gänzlich anderes Prinzip liegt der akustischen Strahlformung zugrunde, welches ein mehrkanaliges Verfahren darstellt [FJZE85]. Dabei wird die Sensitivität einer Mikrophongruppe bezüglich der Sprecherrichtung erhöht, indem ein Sensitivitätsstrahl in diese Richtung ausgebildet wird. Das hat zur Folge, dass Reflexionen des Quellsprachsignals, welche aus anderen als der Sprecherrichtung auf das Mikrophon einfallen, unterdrückt werden, wodurch ein gewisser Enthallungseffekt auftritt. Zusätzlich werden dadurch auch Hintergrundstörungen gedämpft. Eine Schwierigkeit im Zusammenhang mit diesem Verfahren ist die robuste automatische Bestimmung der Sprecherrichtung.

Weiterhin existieren Verfahren, welche die Enthallung mit Hilfe von homomorphischer Entfaltung [SCI75, SPW96] durchführen. Sie sind vom Ansatz her sehr ähnlich zur später in Kap. 3.2.1 vorgestellten Mittelwertsubtraktion und werden daher hier nicht weiter beschrieben.

Abschließend sei noch erwähnt, dass im Prinzip unterschiedliche Kombinationen von Ansätzen vorstellbar sind. So wird zum Beispiel in [KNM06] die Energie des späten Nachhalls mit Hilfe der mehrstufigen linearen Prädiktion geschätzt, um den späten Nachhall durch die Anwendung von spektraler Subtraktion zu unterdrücken. Außerdem können Verfahren zur Enthallung mit Verfahren zur Entstörung wie in [YNM09] verknüpft werden.

3.2. Verfahren zur Extraktion hallrobuster Merkmale

Zu dieser Kategorie gehören Verfahren, welche das Ziel verfolgen, die Merkmalsextraktion derart zu gestalten, dass diese insensitiv gegenüber dem Einfluss von Nachhall und Hintergrundstörungen ist. Darunter befinden sich unter anderem zahlreiche Normierungsansätze sowie auch Methoden, welche sich an der menschlichen Wahrnehmung orientieren. Da das in dieser Dissertation vorgestellte Verfahren ebenfalls in diese Kategorie fällt, wird im Folgenden ein sehr detaillierter Überblick über die merkmalsbasierten Ansätze gegeben.

3.2.1. Normierungsverfahren

Den Normierungsverfahren liegt die Motivation zugrunde, die langzeitigen statistischen Eigenschaften der akustischen Merkmale zu betrachten. Sie gehen von der Feststellung aus, dass sich diese Eigenschaften in Abhängigkeit der Präsenz von Nachhall und Hintergrundstörungen verändern. Als Folge dessen kann eine statistische Fehlanpassung bei der Beschreibung von verhallten und gestörten Sprachsignalen durch das akustische Modell des Spracherkenners auftreten, wenn zuvor das Training unter Verwendung von sauberen Sprachsignalen erfolgt ist. Zur Behebung dieser Diskrepanz lassen sich daher unterschiedliche Normalisierungsstrategien verfolgen, die sich hauptsächlich in der Ordnung und Anzahl der normierten Momente unterscheiden. Dazu muss streng genommen vorausgesetzt werden, dass die entsprechenden Momente überhaupt existieren.

Die praktische Durchführung der Normalisierung erfordert in der Regel eine vorhergehende Schätzung der entsprechenden Momente mit Hilfe der beobachteten Merkmale. Um eine gewisse Genauigkeit dieser Schätzung zu erzielen, muss die Anzahl der dazu herangezogenen Merkmale entsprechend groß sein, wobei sie im Allgemeinen mit der Ordnung des zu schätzenden Momentes steigt. Da die Normierung erst nach der Schätzung stattfinden kann, wird dadurch eine gewisse, oft beträchtliche, Zeitverzögerung im Gesamtsystem eingeführt, worin ein entscheidender Nachteil der Normierungsverfahren liegt. Als Kompromiss lassen sich die Momente mit gleitenden Fenstern schätzen, woran jedoch die Genauigkeit der Schätzung und damit verbunden auch die Effektivität der Normalisierung leidet.

Im Folgenden werden ausgewählte Normierungsverfahren im Detail vorgestellt.

Cepstrale Mittelwertsubtraktion

Der wohl berühmteste Vertreter der Normierungsverfahren ist die sogenannte cepstrale Mittelwertsubtraktion (engl. ***C**epstral **M**ean **S**ubtraction* (*CMS*)) [RLS94], die auf der folgenden Idee basiert. Die zeitdiskrete FOURIER-Transformierte (engl. ***D**iscrete-**T**ime* **F**OURIER ***T**ransform* (*DTFT*)) $S\left(e^{j\theta}\right)$ des verhallten Sprachsignals $s(l)$ lässt sich bekanntlich als Produkt der *DTFT* $X\left(e^{j\theta}\right)$ des sauberen Sprachsignals und der *DTFT* $H\left(e^{j\theta}\right)$ der RIA ausdrücken:

$$S\left(e^{j\theta}\right) = X\left(e^{j\theta}\right) H\left(e^{j\theta}\right). \tag{3.1}$$

In Anlehnung daran lässt sich das Kurzzeit-Spektrum des verhallten Sprachsignals gemäß

$$S(m,k) \approx X(m,k)H(0,k) \tag{3.2}$$

approximieren, falls die zeitliche Ausdehnung des Analysefensters deutlich größer als die der RIA ist [AC07a]. Für den natürlichen Logarithmus des Kurzzeit-Leistungsspektrums gilt dann entsprechend folgende Näherung

$$\ln|S(m,k)|^2 \approx \ln|X(m,k)|^2 + \ln|H(0,k)|^2. \tag{3.3}$$

Alternativ lässt sich eine Approximation direkt im log-MEL-spektralen Bereich gemäß

$$s_{m,q}^{(\mathrm{s})} = \ln\left\{\sum_{k=K_q^{(\mathrm{u})}}^{K_q^{(\mathrm{o})}} |S(m,k)|^2 \Lambda_q(k)\right\} \tag{3.4}$$

$$\approx \ln\left\{\sum_{k=K_q^{(\mathrm{u})}}^{K_q^{(\mathrm{o})}} |X(m,k)|^2 |H(0,k)|^2 \Lambda_q(k)\right\} \tag{3.5}$$

$$\approx \ln\left\{\sum_{k=K_q^{(\mathrm{u})}}^{K_q^{(\mathrm{o})}} |X(m,k)|^2 \Lambda_q(k) \left(\frac{1}{K_q^{(\mathrm{o})} - K_q^{(\mathrm{u})} + 1} \sum_{k=K_q^{(\mathrm{u})}}^{K_q^{(\mathrm{o})}} |H(0,k)|^2\right)\right\} \tag{3.6}$$

$$= x_{m,q}^{(\mathrm{s})} + \ln\left\{\frac{1}{K_q^{(\mathrm{o})} - K_q^{(\mathrm{u})} + 1} \sum_{k=K_q^{(\mathrm{u})}}^{K_q^{(\mathrm{o})}} |H(0,k)|^2\right\} \tag{3.7}$$

angeben, woraus sich unmittelbar ein analoger Ausdruck im Cepstrum gewinnen lässt. Dabei ist zu berücksichtigen, dass die jeweils letzten Terme in (3.3) und (3.7) segmentunabhängig sind. Subtrahiert man folglich vom logarithmischen Kurzzeit-Leistungsspektrum oder vom Cepstrum eines verhallten Sprachsignals seinen Mittelwert, so wird der Einfluss der RIA näherungsweise eliminiert. Da das logarithmische Kurzzeit-Leistungsspektrum und das Cepstrum gewöhnlich nicht mittelwertfrei sind, muss die Subtraktion ebenfalls bei der Extraktion der Merkmale für das Training des Spracherkenners stattfinden.

Nun beträgt die Dauer eines Analysefensters zur Merkmalsextraktion in der Regel etwa 25 ms. Hingegen ist die zeitliche Ausdehnung einer typischen Impulsantwort deutlich länger und liegt im Bereich von einigen Hundert Millisekunden. Daher ist *CMS* in der oben beschriebenen Form nicht dazu in der Lage, den Einfluss von Nachhall auf der Merkmalsebene zu reduzieren, worin eine wesentliche Schwachstelle dieses Ansatzes besteht. Er eignet sich viel eher dazu, den bei der Aufnahme der Sprachsignale durch Mikrophone mit unterschiedlichen Frequenzcharakteristiken entstehenden Auswirkungen zu unterdrücken [RLS94].

Als Abhilfe wurde in [ATH97] vorgeschlagen, deutlich längere Analysefenster der Dauer von etwa 2 s für die cepstrale Mittelwertsubtraktion zu verwenden. Um die resultierenden Merkmale für die Spracherkennung nutzen zu können, müssen diese wieder in das gewöhnliche Format umgerechnet werden. Damit ist gemeint, dass die für den Erkenner übliche Zeit-Frequenz-Auflösung wiederhergestellt werden muss. Dazu wird in [Ave97] eine approximative Transformation des Kurzzeit-Leistungsspektrums hergeleitet, welche dessen zeitliche Auflösung zulasten der Frequenzauflösung vergrößert. Anstatt diese sogenannte partielle Synthese vorzunehmen, ist es auch möglich, das akustische Signal nach der Anwendung von *CMS* zu resynthetisieren, um anschließend eine gewöhnliche Merkmalsextraktion durchzuführen [GM01]. Für diesen als Langzeit-*CMS* bezeichneten Ansatz wird zur Resynthese neben dem Kurzzeit-Leistungsspektrum im Grunde noch die Kurzzeit-Phase des sauberen Sprachsignals benötigt. Da sie jedoch im Allgemeinen unbekannt ist, wird statt dessen die Kurzzeit-Phase des verhallten Signals verwendet. Mit Hilfe eines derartigen Verfahrens wird zwischenzeitlich ein enthalltes Sprachsignal berechnet, weshalb es eigentlich zu den signalbasierten Ansätzen gehört. Es konnte damit eine beachtliche Reduktion der Wortfehlerrate in Gegenwart von sowohl künstlichem als auch natürlichen Nachhall im Vergleich zur Merkmalsextraktion gemäß dem *ETSI*-Standard [ETSb] erzielt werden [GM01]. Da die Mittelung

über 21 aufeinander folgende Segmente mit einem Überlapp von 50 % ausgeführt wurde, entsprach die durch das Verfahren eingeführte zeitliche Verzögerung etwa 11 s.

Weiterhin lässt sich Langzeit-*CMS* beispielsweise mit spektraler Subtraktion kombinieren, um eine gemeinsame Enthallung und Entstörung akustischer Merkmale vorzunehmen [GM02]. Für die spektrale Subtraktion kann ein zeitlich konstantes Kurzzeit-Leistungsspektrum der Hintergrundstörung angenommen werden, welches für die Dauer einer Sprachäußerung gültig ist und mit Hilfe einer Sprachaktivitätsdetektion (engl. ***V**oice **A**ctivity **D**etection* (*VAD*)) geschätzt wird.

Cepstrale Varianznormierung

Die cepstrale Varianznormierung (engl. ***C**epstral **V**ariance **N**ormalization* (*CVN*)) wurde bedingt durch die historische Entwicklung der Spracherkennung zunächst im Sinne der Kompensation von Hintergrundstörungen eingesetzt [CB07, VL98]. Die ursprüngliche Motivation für ihre Anwendung lag in der Beobachtung, dass aufgrund der Hintergrundstörungen energiearme Bereiche des Kurzzeit-Leistungsspektrum ”aufgefüllt“ werden, sodass sich in letzter Konsequenz die Varianz einzelner cepstraler Merkmale reduziert. Ein ähnlicher Effekt tritt jedoch bedingt durch den zeitlich dispersiven Effekt des Nachhalls auf, so dass *CVN* dazu in der Lage ist, die Robustheit der Merkmalsextraktion gegenüber Nachhall in gewissem Maße zu steigern [TTN07].

Histogrammangleichung

Der Grenzfall der Normierung einzelner Merkmalsvektorkomponenten bezüglich aller ihrer Momente kann äquivalent als gezielte Angleichung ihrer Verteilungsdichtefunktion an eine Referenz angesehen werden [dlTPS$^+$05, TTN07]. Dabei wird implizit vorausgesetzt, dass zeitlich aufeinander folgende Merkmalsvektorkomponenten Realisierungen unabhängiger und identisch verteilter Zufallsvariablen darstellen. Insbesondere folgt aus einer solchen Voraussetzung, dass die zeitliche Trajektorie einer Merkmalsvektorkomponente als Realisierung eines stationären Prozesses interpretiert werden kann, was im Falle von zugrunde liegenden Sprachsignalen eigentlich nicht sinnvoll ist.

Die Notwendigkeit einer Angleichung der Verteilungsdichtefunktion erwächst nun basierend auf dieser Annahme dadurch, dass der gemeinsame Effekt von Nachhall und Hintergrundstörungen approximativ zu einer nichtlinearen Transformation der cepstralen Merkmale führt. Dazu muss zunächst die Verteilungsdichtefunktion einzelner Merkmalsvektorkomponenten durch ein empirisch bestimmtes Histogramm hinreichend genau approximiert werden, was offensichtlich eine ausreichende Menge an Beobachtungen erfordert. Um einen Kompromiss zwischen der Zeitverzögerung des Verfahrens und einer möglichst großen Genauigkeit zu erreichen, wird die Schätzung des Histogramms gewöhnlich auf der Grundlage ganzer Sprachäußerungen durchgeführt [TTN07]. Anschließend wird jede Merkmalsvektorkomponente derart transformiert, dass das resultierende Histogramm einem Referenzhistogramm entspricht. Man spricht in der Literatur deshalb auch von einer sogenannten Histogrammangleichung. Dabei muss angenommen werden, dass die entsprechende Transformation existiert, was äquivalent dadurch ausgedrückt werden kann, dass die Verteilungsfunktion streng monoton wachsend ist. In der Praxis wird man sich damit begnügen, dass diese Eigenschaft nur näherungsweise erfüllt ist, da ihre Verifikation aufgrund von Fehlern bei der

empirischen Schätzung der Verteilungsfunktion mit Hilfe von normierten kumulativen Histogrammen unmöglich ist.

Das Verfahren bietet zwei wesentliche Vorteile. Zum einen lässt sich der Rechenaufwand relativ gering halten, indem die gewöhnlich nichtlineare Transformation der Merkmalsvektorkomponenten mit Hilfe von Nachschlagetabellen realisiert wird. Zum anderen werden keine Annahmen über die Art der Transformation getroffen, so dass sich die Methode prinzipiell zur Kompensation unterschiedlichster Arten von Störungen eignet. Es muss jedoch betont werden, dass besonders im Falle von Nachhall eine starke Abhängigkeit zwischen zeitlich aufeinander folgenden Merkmalsvektorkomponenten besteht, welche der bereits erwähnten Unabhängigkeitsannahme des Verfahrens deutlich widerspricht und folglich seine Effektivität enorm verringert. Nichtsdestotrotz konnte mit der auf das Cepstrum angewendeten Histogrammangleichung eine merkbare Steigerung der Leistungsfähigkeit des Spracherkenners in Gegenwart von künstlich erzeugtem Nachhall erzielt werden [TTN07].

Affine Transformation von Merkmalsvektoren

Bei den bisher vorgestellten Normalisierungsstrategien wurden einzelne Merkmalsvektorkomponenten getrennt voneinander betrachtet. Nun ist es jedoch auch möglich, eine affin lineare Transformation auf den gesamten Merkmalsvektor anzuwenden, wobei das Kriterium zur Bestimmung der Transformation in der Maximierung der sogenannten Likelihoodfunktion für eine Menge von Adaptionsdaten liegt. Dieser Ansatz wird im Englischen als ***C**onstrained **M**aximum **L**ikelihood **L**inear **R**egression* (*CMLLR*) oder alternativ als ***F**eature-space **M**aximum **L**ikelihood **L**inear **R**egression* (*FMLLR*) bezeichnet [Gal98].

In der Regel können die Auswirkungen des Nachhalls auf das Cepstrum unter der Annahme von gewöhnlichen Analysefensterlängen nicht durch affin lineare Transformationen ausgedrückt werden. Denn die zeitliche Verschmierung des Cepstrums erzeugt eine starke Abhängigkeit von aufeinander folgenden Merkmalsvektoren. Beinhalten die Merkmalsvektoren jedoch dynamische Komponenten, welche diese Abhängigkeit in einer gewissen Weise erfassen, lässt sich die Anwendung von *CMLLR* zur Robustheit gegenüber Nachhall zumindest in Ansätzen rechtfertigen. So wurde es vom Autor dieser Dissertation bereits in [KHU10] durchaus erfolgreich zur Merkmalsenthallung eingesetzt. Die Resultate hingen jedoch stark von der Menge der Adaptionsdaten sowie davon ab, ob deren Transkription zur Bestimmung der Transformation vorlag.

3.2.2. Perzeptuell motivierte Verfahren

Die in diesem Abschnitt vorgestellten Verfahren gehen im Wesentlichen von der grundlegenden Feststellung aus, dass die Aufgabe der Spracherkennung in der Dekodierung einer linguistischen Nachricht liegt, welche ursprünglich durch den Menschen beim Sprechen in die Bewegungen des Vokaltraktes codiert wurde [HMBK91]. Da die physikalischen Eigenschaften des Vokaltraktes, vor allem seine Trägheit, nur gewisse Änderungsraten seiner Stellung zulassen, prägen sie dadurch die Eigenschaften eines Sprachsignals. Diese Tatsache lässt sich demzufolge auch bei der perzeptuell orientierten Analyse eines akustischen Signals innerhalb einzelner kritischer (Frequenz-)Bänder beobachten. Fasst man nämlich die entsprechenden Bandpasssignale approximativ als amplitudenmodulierte Signale auf, so besitzen die zugehörigen Einhüllenden hauptsächlich Anteile für Modulationsfrequenzen im Bereich

zwischen 0,5 Hz und 16 Hz [HS85]. Insbesondere ist in diesem Zusammenhang bemerkenswert, dass das menschliche Gehör gegenüber Modulationsfrequenzen im Bereich von etwa 4 Hz eine erhöhte Sensitivität aufweist [HM94], welche der Rate von Silben innerhalb der Sprache [HSP80] entspricht.

Berechnung relativer Kurzzeit-Leistungspektren

Die auf relativen Kurzzeit-Leistungsspektren basierenden Merkmale (engl. ***Rel**ative **Spec**tr**a**l* (*RASTA*) *features*) [HMBK91, HM94] basieren ursprünglich auf der Beobachtung, dass für die menschliche Wahrnehmung hauptsächlich relative Unterschiede der Stimulation von Bedeutung sind. Orientierend daran wird deshalb in [HMBK91] vorgeschlagen, eine Abkehr von der bis dahin etablierten Verwendung absoluter Werte des Kurzzeit-Leistungsspektrums zur Merkmalsextraktion vorzunehmen.

Die *RASTA*-Merkmale stellen eine modifizierte Version von Merkmalen dar, welche auf einer perzeptuell motivierten linearen Prädiktion (engl. ***P**erceptual **L**inear **P**rediction* (*PLP*)) basieren [HHW85, Her90]. Für die Berechnung der *PLP*-Merkmale wird in einem ersten Schritt die Leistung des Sprachsignals innerhalb der einzelnen kritischer Bänder [Gre61] bestimmt. Dieses geschieht unter Verwendung des Kurzzeit-Leistungsspektrums auf eine ähnliche Weise wie für die Berechnung der *MFCC*s in (2.4). Der einzige Unterschied liegt in der Verwendung von Fensterfunktionen, welche bezüglich der MEL-Frequenzskala eine trapezförmige Gestalt aufweisen. Im Anschluss erfolgt eine Gewichtung sowie Komprimierung der Leistung innerhalb der kritischen Bänder zur approximativen Nachahmung der perzeptuellen Lautheit. Das resultierende verzerrte Kurzzeit-Leistungsspektrum wird dann durch ein autoregressives Modell approximiert, indem *LPC*-Koeffizienten berechnet werden. Der aus den *LPC*-Koeffizienten bestehende Vektor wird anschließend ins Cepstrum transformiert.

Die Modifikation der *PLP*-Merkmale besteht nun in der Einführung einer kompressiven Nichtlinearität, einer Bandpass-Filterung sowie einer dekompressiven Nichtlinearität nach der Berechnung des Kurzzeit-Leistungsspektrums für kritische Bänder [HMBK91]. Die grundsätzliche Idee der Bandpass-Filterung besteht in der Unterdrückung aller besonders schnell oder besonders langsam veränderlichen Komponenten in der zeitlichen Trajektorie der komprimierten Leistung einzelner kritischer Bänder, da diese typischerweise nicht die linguistische Nachricht enthalten. So ähnelt der Durchlassbereich des Bandpass-Filters dem bereits zu Beginn von Kap. 3.2.2 erwähnten Frequenzbereich zwischen 0,5 Hz und 16 Hz.

Variationen des Verfahrens entstehen beispielsweise durch unterschiedliche Wahlen der kompressive Nichtlinearität. So zeichnen sich die sogenannten LOG-*RASTA-PLP*-Koeffizienten [HMBK91] durch eine logarithmische Kompression aus, welche besonders geeignet ist, um Effekte von Faltungsstörungen zu unterdrücken und damit eine Robustheit gegenüber Kanaleinflüssen zu erzielen. Dabei wird dasselbe Prinzip der Additivität der Faltungsstörung im logarithmischen Kurzzeit-Leistungsspektrum wie auch bei der cepstralen Subtraktion ausgenutzt. Die sogenannten LIN-LOG-*RASTA-PLP*-Koeffizienten nutzen eine kompressive Linearität, welche approximativ linear für kleine Werte des Argumentes und approximativ logarithmisch für große Werte des Argumentes ist, wobei die Grenze zwischen den beiden Bereichen signalabhängig gewählt wird [HM94]. Damit lassen sich zusätzlich zu Faltungsstörungen additive Hintergrundstörungen unterdrücken, welche approximativ additiv im linearen Kurzzeit-Leistungsspektrum sind.

In experimentellen Untersuchungen wurde festgestellt, dass die alleinige Verwendung von

RASTA-PLP-Koeffizienten im Vergleich zur Verwendung der *PLP*-Koeffizienten zu keiner Leistungssteigerung des Spracherkenners in Gegenwart von Nachhall führte [KM97]. Dieses änderte sich jedoch, als für die Spracherkennung beide Arten von Koeffizienten gemeinsam verwendet wurden. Bei diesem Ansatz besteht eine starke Parallele zur Ergänzung der *MFCC*s durch die DELTA-Merkmale zur Erfassung eines gewissen zeitlichen Kontexts (siehe auch Kap. 2.1). Die Berechnung der DELTA-Merkmale kann als ein Spezialfall der *RASTA*-Verarbeitung aufgefasst werden kann, wobei die Bandpassfilterung mit Hilfe eines nicht kausalen Filters mit endlicher Impulsantwort vorgenommen wird [HM94]. An dieser Stelle soll darauf hingewiesen werden, dass auch die cepstrale Mittelwertsubtraktion eine große Ähnlichkeit zur *RASTA*-Verarbeitung aufweist, wobei jedoch die Bandpassfilterung durch eine Hochpassfilterung zur ausschließlichen Entfernung des Gleichanteils ersetzt ist.

Weiterhin existieren Ansätze für den Entwurf von datenabhängigen Bandpass-Filtern mit Hilfe der linearen Diskriminantenanalyse unter Verwendung von verhallten Testsprachsignalen [vVH97]. Dabei findet im Wesentlichen eine Anpassung des Durchlassbereiches an das Ausmaß des Nachhalls statt. Die Verwendung derartiger Methoden in Gegenwart von Nachhall offenbarte jedoch eine starke Sensibilität des Verfahrens im Bezug auf die Wahl von Trainingsdaten [SC00], wobei bei einer Fehlanpassung der Trainingsdaten an die Testdaten die Erkennungsleistungen sehr schlecht ausfallen können.

Modulationsspektrogramm

Das Modulationsspektrogramm stellt eine Verallgemeinerung der *RASTA-PLP*-Algorithmen dar [GK97] [KMG98]. Das Sprachsignal wird hierbei auch in Anlehnung an Studien zur menschlichen Wahrnehmung in kritischen Bändern analysiert, wobei jedoch anstelle einer Kurzzeit-Spektralanalyse mittels der *DFT* eine Bank von Bandpass-Filtern mit endlicher Impulsantwort zum Einsatz kommt. Die Bandpasssignale werden abschnittsweise approximativ als amplitudenmodulierte Signale aufgefasst, wobei das Ziel in der Darstellung von Amplitudenmodulationen in ihrer Stärke und zeitlichem Verlauf im Bereich zwischen 0 Hz und 8 Hz mit einer besonders hohen Sensitivität bei 4 Hz besteht. Dazu wird die Einhüllende der Bandpasssignale bestimmt und zunächst einer Energienormalisierung unterworfen. Anschließend erfolgt eine Bandpass-Filterung der Einhüllenden, wobei die Impulsantwort des Bandpasses ein HAMMING-Fenster darstellt, welche durch eine komplexe Exponentialschwingung der Frequenz von 4 Hz moduliert wird. Die Wirkung dieser Operation ähnelt der eines signalangepassten Filters (engl. *matched filter*) zur Detektion von Signalen mit einer temporalen Struktur, die derjenigen der Sprache entspricht (siehe Bemerkungen zu Beginn von Kap. 3.2.2) [KMG98]. Als Folge dessen fällt die meiste Energie im Modulationsspektrogramm auf den Bereich von silbischen Kernen. In experimentellen Untersuchungen hat sich gezeigt, dass das Modulationsspektrogramm bei Präsenz von gemäßigtem Nachhall keine Vorteile gegenüber den *RASTA*-Merkmalen im Hinblick auf die Spracherkennung bringt [KMG98]. Hingegen konnten durch eine Kombination beider Methoden Verbesserungen gegenüber der alleinigen Verwendung der *RASTA*-Merkmale erzielt werden.

Analyse innerhalb Teilbändern mit linearer Prädiktion im Frequenzbereich

Ähnlich wie beim Modulationsspektrogramm wird in [TGH08a] die Einhüllende von Teilbandsignalen betrachtet und innerhalb sich nicht überlappender Segmente der Dauer von

etwa 1 s analysiert. Die Untersuchung vollzieht sich jedoch mit Hilfe von linearer Prädiktion im Frequenzbereich (engl. ***F**requency **D**omain **L**inear **P**rediction* (*FDLP*)), wobei eine geglättete, minimalphasige, parametrische Darstellung der zeitlichen Einhüllenden berechnet wird. Die Methode orientiert sich an dem Vorbild der linearen Prädiktionscodierung [RJ93, Kap. 3.3] [Mak75], wobei autoregressive Modelle zur parametrischen Repräsentation der spektralen statt der zeitlichen Einhüllenden genutzt werden.

Ein wesentlicher Aspekt im Zusammenhang mit der Erkennung verhallter Sprache ist bei diesem Ansatz die Tatsache, dass sich die spektrale Autokorrelationsfunktion eines zu einem verhallten Sprachsignal zugehörigen Teilbandsignals approximativ als Produkt zweier weiterer Autokorrelationsfunktionen ausdrücken lässt, nämlich der des entsprechenden Teilbandsignals zugehörig zum sauberen Sprachsignal sowie der des Teilbandsignals zugehörig zur RIA [TGH08b]. Die Herleitung dieser Aussage stützt sich darauf, dass zwischen der komplexen Einhüllenden des verhallten Sprachsignals, des sauberen Sprachsignals und der RIA ein Zusammenhang besteht, der sich näherungsweise durch eine Faltung beschreiben lässt [MH83]. Da die komplexe Einhüllende eines Bandpasssignals die inverse FOURIER-Transformierte dessen spektraler Autokorrelationsfunktion bildet [Her96], lässt sich die Aussage über die Multiplikativität der spektralen Autokorrelationsfunktionen durch Ausnutzung der Dualität zwischen dem Zeit- und Frequenzbereich gewinnen. Nimmt man nun weiter an, dass für die RIA die spektrale Autokorrelationsfunktion einzelner Teilbandsignale nur sehr langsam ändert, lässt sich durch eine bandspezifische Amplitudennormierung der komplexen Einhüllenden der Einfluss der RIA unterdrücken.

Nach der Anwendung der linearen Prädiktion im Frequenzbereich und der Normierung erhält man eine Menge von Einhüllenden für einzelne Teilbänder, welche als Zeit-Frequenz-Repräsentation angesehen werden können. Diese wird anschließend bezüglich der Zeit auf 100 Hz unterabgetastet, um eine gewisse Konformität mit der gewöhnlichen Zeit-Frequenz-Auflösung bei der Merkmalsextraktion herzustellen. Die resultierenden Kurzzeit-Energien einzelner Segmente zusammengefasst über alle Subbänder werden danach ins Cepstrum transformiert.

Ergebnisse der in [TGH08b] durchgeführten experimentellen Untersuchungen zeigen beispielsweise einen deutlichen Vorteil des Verfahrens gegenüber *CMS* und Langzeit-*CMS* im Bezug auf die Erkennung von verhallter Sprache. Die Leistungsfähigkeit kann dabei enorm durch die Vergrößerung der Segmentlänge und der Vergrößerung der spektralen Auflösung gesteigert werden, wodurch die Multiplikativitätsaussage bezüglich der spektralen Autokorrelationsfunktion in ihrer Güte verbessert und somit die Normalisierung effektiver wird.

Modulationsanalyse als Ergänzung der *MFCC*s

In [MM10] wurde vorgeschlagen, eine abgewandelte Form der *MFCC*s mit Hilfe von auf der Modulationsanalyse beruhenden Koeffizienten zu ergänzen. Die Modifikation der *MFCC*s besteht prinzipiell in der Verwendung einer sogenannten GAMMATONE-Filterbank zur Extraktion der Signale für einzelne kritische Bänder anstatt der Durchführung einer MEL-Filterung beruhend auf dem Kurzzeit-Leistungsspektrum. Bei GAMMATONE-Filtern handelt es sich um lineare Filter, welche die physiologisch motivierten Verarbeitung durch die Cochlea nachahmen [PRH$^+$92].

An Stelle dynamischer Merkmale wie der DELTA-Merkmale, welche die zeitliche Entwicklung der *MFCC*s beschreiben, werden Merkmale verwendet, welche die Energie von

Modulationen im Frequenzbereich zwischen 2 Hz und 16 Hz in der Trajektorie einzelner cepstraler Koeffizienten darstellen. Das Modulationsspektrum wird mit Hilfe der FOURIER-Transformation der zeitlichen Trajektorie der Energien innerhalb von Teilbändern berechnet. Als Merkmal wird die Energie im Frequenzband zwischen 2 Hz und 16 Hz verwendet. Anschließend werden die DELTA-DELTA-Merkmale durch numerische Differenziation der Merkmale beruhend auf der Modulationsanalyse bestimmt.

Bezüglich der experimentellen Ergebnisse lässt sich zusammenfassen, dass bereits durch die Ersetzung der MEL-spektralen Koeffizienten durch die auf der GAMMATONE-Filterung basierenden Merkmale die Erkennungsleistung in Gegenwart von Nachhall deutlich gesteigert werden konnte. Durch den Austausch der DELTA-Merkmale konnte eine weitere Verbesserung der Erkennungsleistung erzielt werden [MM10]. Ein wahrscheinlicher Grund dafür liegt darin, dass man sich bei der Erfassung der zeitlichen Veränderungen durch die alternativen dynamischen Merkmale auf den linguistisch relevanten Frequenzbereich konzentriert.

3.2.3. Sonstige merkmalsbasierte Verfahren

Abgesehen von den merkmalsbasierten Verfahren der ersten beiden Kategorien existieren in der Literatur weitere Ansätze, die in dieser Dissertation nicht in der ganzen Ausführlichkeit vorgestellt werden können. Es werden daher nur einige ausgewählte Verfahren kurz erläutert.

Berechnung der dynamischen Merkmale auf Grundlage der linear skalierten Energie

In [IFN10] wird vorgeschlagen, die Berechnung der dynamischen Merkmale im linearen statt dem gewöhnlichen logarithmischen Energiebereich durchzuführen. Sie werden motiviert durch die Tatsache, dass die Energie des Nachhalls einen exponentiellen Abklang aufweist, welcher jedoch durch die Anwendung des Logarithmus affin linear wird. Als Folge dessen bleiben die Werte der dynamischen Merkmale vorwiegend in kurzen Sprachpausen lange unerwünscht konstant, so dass der Spracherkenner nicht vorhandene Wörter erkennt. Da der Dynamikbereich der (linearen) Energie deutlich größer ist und die Energie des Nachhalls exponentiell, also inbesondere sehr schnell, abklingt, werden die dynamischen Merkmale deutlich weniger durch den Nachhall gestört. Um sicherzustellen, dass die Merkmale eine approximativ GAUSS-förmige Verteilungsdichtefunktion besitzen, welche für eine Modellierung mit Hilfe von HMMs im Spracherkenner notwendig ist, muss zusätzlich eine geeignete Normierung vorgenommen werden.

Ausnutzung der harmonischen Struktur der Sprache

Eine weitere Art zur Extraktion robuster Merkmale geht von der Annahme aus, dass harmonische Komponenten der Sprache durch den Nachhall nur geringfügig verändert werden [PLLH08]. Werden sie jedoch von stimmlosen Lauten gefolgt, werden diese durch die abklingende Energie der stimmhaften Laute überlagert. Der Einfluss auf die stimmlosen Laute ist besonders groß im niederfrequenten Bereich, wo die stimmlosen Laute in der Regel wenig Energie besitzen. Folglich besteht die Idee in [PLLH08] unter anderem darin, stimmhafte und stimmlose Bereiche innerhalb des Sprachsignals zu detektieren und jegliche Energie in unteren Teilbändern innerhalb von stimmlosen Bereichen zu entfernen. Dieser Ansatz wurde weiterhin in [PLU+08] mit einer darauf folgenden Analyse des Modulationsspektrums

kombiniert.

Merkmalsverbesserung

Ähnlich dem in dieser Dissertation verfolgten Ansatz wird in [Wöl09] die gemeinsame Enthallung und Entstörung der Merkmale als ein Problem der Verfolgung einer Trajektorie aufgefasst, welches mit Hilfe einer BAYES'schen Methode gelöst wird. Ein entscheidender Unterschied zum Verfahren, welches in dieser Dissertation vorgeschlagen wird, liegt dabei darin, dass der Nachhall als additive Störung im MEL-spektralen Bereich aufgefasst wird, dessen Ausmaß zunächst im Zeitbereich mit Hilfe der mehrstufigen linearen Prädiktion [GD97] geschätzt wird. Weitere deutliche Unterschiede bestehen in den verwendeten A-priori-Modelle zur Beschreibung der Sprache und Störung im Merkmalsraum sowie der Realisierung der Inferenz, wozu eine Partikelfilterung genutzt wird.

Ausnutzung von Unsicherheitsinformation

Die in [PBB02] präsentierte Methode wird motiviert durch die Feststellung, dass das menschliche Gehörsystem einen Mechanismus besitzt, um mit unverlässlichen "Daten" umzugehen [CGJV01]. Demzufolge wird versucht, verlässliche Bereiche im Kurzzeit-Leistungsspektrum aufzufinden, um diese anschließend an einen modifizierten Spracherkenner weiterzuleiten. Insofern handelt es sich bei diesem Ansatz um eine Kombination aus einer merkmalsbasierten und modellbasierten Methode, was als Nachteil bedingt durch den erforderlichen Eingriff in den Erkenner gesehen werden kann.

Verlässliche Bereiche im Kurzzeit-Leistungsspektrum sind in der Regel dadurch gekennzeichnet, dass sie energiereich sind und dadurch nicht stark durch den Einfluss des Nachhalls verändert werden. Zum Auffinden dieser wird eine sogenannte Hallmaske verwendet. Damit ein Bereich als verlässlich gilt, muss seine Energie eine vorgegebene, zuvor empirisch ermittelte Schranke übersteigen.

3.3. Verfahren basierend auf der Modifikation des akustischen Modells oder des Decoders

Eine weitere Möglichkeit zur Kompensation der Auswirkungen des Nachhalls auf die statistischen Eigenschaften der akustischen Merkmale besteht in der Modifikation des akustischen Modells oder des Decoders.

3.3.1. Modifikation des akustischen Modells

Das akustische Modell lässt sich beispielsweise dadurch modifizieren, dass ein auf das Erkennungsszenario angepasstes Training mit verhallten und eventuell zusätzlich gestörten Sprachsignalen durchgeführt wird. Da jedoch das Erkennungsszenario oft zum Zeitpunkt des Trainings noch unbekannt ist, wird eine vielfältige und umfangreiche Menge an Trainingsdaten benötigt, um vorab möglichst viele Einsatzbedingungen abzudecken. An Stelle der Verwendung von echten Sprachäußerungen, deren Aufnahme aufwendig und teuer ist,

bietet sich eine künstliche, modellbasierte Erzeugung der Daten [GMOS99, SFB01] an. Dabei ist es sinnvoll, eine Parametrisierung der künstlich erzeugten Sprachdaten und der damit trainierten akustischen Modelle mit Hilfe der Nachhallzeit T_{60} vornehmen. Zur Spracherkennung muss anschließend nur noch das passende akustische Modell beruhend auf einer Schätzung der Nachhallzeit ausgewählt werden [CC04]. Der Nachteil eines solchen Verfahrens liegt in der großen Datenmenge, die zur Erfassung sämtlicher akustischer Modelle notwendig ist.

Eine dazu alternative Methode ist die Adaption von akustischen Modellen, welche mit sauberen Sprachsignalen trainiert wurden. Dabei unterscheidet man grundsätzlich zwischen der statischen und der dynamischen Adaption.

Bei der statischen Adaption werden die akustischen Modelle vorab einmal an das Erkennungsszenario angepasst und bei der Erkennung nicht mehr verändert. Ein in diesem Zusammenhang zu nennender Ansatz ist die Anwendung von affin linearen Transformationen auf einzelne Komponenten der *GMM*s zur Modellierung der Emissionsverteilungsdichtefunktionen von HMM-Zuständen. Da das Kriterium zur Bestimmung der Transformation die Maximierung der Likelihood beruhend auf einer gegebenen Menge von Adaptionsdaten ist, wird diese Methode im Englischen als ***M**aximum **L**ikelihood **L**inear **R**egression* (*MLLR*) bezeichnet [GW96, Gal98]. Sie unterscheidet sich vom in Kap. 3.2.1 vorgestellten *CMLLR* dadurch, dass statt einer globalen Transformation für alle Emissionsverteilungsdichtefunktionen nun viele unterschiedliche Transformationen abhängig vom HMM-Zustand und *GMM*-Komponente ermöglicht werden. Die Menge der verschiedenen Transformation lässt sich im Prinzip durch die Menge der zur Verfügung stehenden Adaptionsdaten steuern, da gleiche Transformationen von vielen HMM-Zuständen und *GMM*-Komponenten gemeinsam geteilt werden können. Dadurch lässt sich eine sinnvolle Adaption des akustischen Modells bereits mit einer geringen Menge an Adaptionsdaten bewerkstelligen. Wie auch *CMLLR* wurde *MLLR* ursprünglich zur Adaption des akustischen Modells an unterschiedliche Sprecher eingeführt. In [TTN06] wurde es jedoch auch zur Kompensation der Effekte des Nachhalls eingesetzt. Die Wirkung von *MLLR* ist dabei hauptsächlich auf das Vorhandensein von dynamischen Komponenten innerhalb des Merkmalsvektors zurückzuführen, wodurch ein gewisser zeitlicher Kontext erfasst wird. Dieser Kontext ist beispielsweise bei der in Kap. 2.1 beschriebenen Merkmalsextraktion auf 6 zeitlich vorhergehende Segmente beschränkt (vgl. dazu Parameter in Tab. 2.1). Geht das Ausmaß der zeitlichen Verschmierung darüber hinaus, so kann der Effekt nicht mehr ausreichend kompensiert werden. Aus dieser Sicht es vernünftig, *MLLR* wie in [MOG00] im Sinne einer inkrementellen Adaption von akustischen Daten, welche bereits mit verhallten Sprachsignalen trainiert wurden, zur Reduktion der verbleibenden Fehlanpassung zu nutzen. Ein weiteres Problem von *MLLR* stellt die Tatsache dar, dass die Transkription des Adaptionsdaten für die Adaption bekannt sein muss. Da diese Voraussetzung gewöhnlich nicht gegeben ist, wird diese durch eine vorhergehende Erkennung mit Hilfe von nicht adaptierten Modellen gewonnen. Man spricht dabei von unüberwachter Adaption, die aufgrund einer gewöhnlich fehlerhaften Transkription die Leistungsfähigkeit der Adaption zusätzlich verringert.

Eine aus physikalischer Sicht genauere Modellierung wird erzielt, indem die Auswirkung des Nachhalls auf das MEL-Spektrum durch eine Faltung dessen bezüglich der Zeit mit einer Funktion, die in engem Zusammenhang mit der RIA steht, approximiert wird (siehe z. B. Kap. 5.2.2 oder auch [SK08]). Diese Beschreibung lässt sich beispielsweise zur Modelladaption durch Aufspaltung von HMM-Zuständen in einzelne Teilzustände verwenden, um

damit unterschiedliche Kompensationen abhängig von der genauen Verweildauer innerhalb eines HMM-Zustandes vornehmen zu können [RNS05c, RNS05b]. Die Anzahl der Teilzustände hängt dann von der mittleren Verweildauer innerhalb eines HMM-Zustandes ab. Die Repräsentation der RIA im MEL-spektralen Bereich lässt sich beispielsweise mit Hilfe von Adaptionsdaten unter Verwendung des akustischen Modells für saubere Sprachsignale durchführen [RNS05b].

Alternativ lässt sich die Modellanpassung durch eine sogenannte parallele Modellkombination (engl. ***P**arallel **M**odel **C**ombination* (*PMC*)) erzielen [GY95]. Diese wurde ursprünglich entwickelt, um akustische Modelle der Sprache und der Hintergrundstörung geeignet zu kombinieren. Sie basiert auf der Annahme, dass die Sprache und die Hintergrundstörung im MEL-Spektrum approximativ additiv sind. In Folge dessen werden dazu die akustischen Modelle vom cepstralen in den MEL-spektralen Bereich transformiert, dort zusammengesetzt und entsprechend zurücktransformiert. Für die Kombination werden nur Modifikationen der ersten beiden Momente einzelner *GMM*-Komponenten für das Cepstrum in Betracht gezogen, weshalb diese relativ einfach vorzunehmen ist. Die Kombination ist jedoch höchst approximativ, da sie unter anderem annimmt, dass die Summe zweier log-normalverteilter Variablen wieder log-normalverteilt ist [GY95].

Unter Ausnutzung der Tatsache, dass die Auswirkungen des Nachhalls durch eine Faltung im MEL-spektralen Merkmalsbereich beschrieben werden können, lässt sich die ursprünglich eingeführte parallele Modellkombination zur entsprechenden Adaption der akustischen Modelle modifizieren [RNS05a, HGH06]. Dabei muss berücksichtigt werden, dass die Adaption auf der Basis von HMM-Zuständen und nicht Merkmalsvektoren erfolgt. Sie bedient sich in [HGH06] eines Modells der *EDC* einer RIA, wobei die *EDC* durch eine exponentiell abklingende Funktion approximiert wird und als einzigen Parameter die Nachhallzeit T_{60} besitzt. Damit kann durch Beachtung der mittleren Verweildauer in einem HMM-Zustand der mittlere Anteil der Energie berechnet werden, welcher auf die zeitlich folgenden HMM-Zustände verschmiert wird.

In [HGH06] wird die Adaption unabhängig auf einzelne HMMs, welche ganze Wörter modellieren, angewendet. Es findet demnach keine Berücksichtigung der Verschmierung der Energie über Wortgrenzen hinweg statt. Dies ist ein Problem, was im Allgemeinen bei der statischen Adaption auftritt. Denn die Energie des Nachhalls hängt in hohem Maße vom vorhergehenden Kontext eines HMM-Zustandes ab, der vor der eigentlichen Erkennung natürlich nicht bekannt ist. Ein gewisser vorhergehender, HMM-übergreifender Kontext kann bei der Adaption von triphonbasierten HMMs, welche zur Erkennung von Sprache mit großem Vokabular eingesetzt werden, genutzt werden [HF08]. Denn ein Triphon beschreibt ein Phonem in Abhängigkeit seines Vorgänger- und Nachfolgephonems. Der Kontext reicht jedoch gewöhnlich nicht aus, um den Ursprung der Energie des Nachhalls ausreichend zu erfassen. Denn die mittlere Dauer eines Phonems beträgt etwa 125 ms [RJ93, Kap. 2] und ist damit deutlich kürzer als die Nachhallzeit, die in gewöhnlichen Büros und Wohnzimmern einige Hundert Millisekunden betragen kann [Kut00].

Die dynamische Adaption der akustischen Modelle an den Nachhall findet parallel zur Dekodierung statt [YNS04, TN04, SMK11]. Sie bietet den großen Vorteil, dass sich durch die Dekodierung ein wahrscheinlicher, vorhergehender Kontext zu einem HMM-Zustand erschließt, wodurch die Energie des Nachhalls deutlich besser modelliert werden kann. Dieser Vorteil wird jedoch zulasten eines deutlich erhöhten Rechenaufwandes bei der Spracherkennung erkauft.

In [TN04] findet eine dynamische Adaption auf der Grundlage eines rekursiven Beobachtungsmodells zur Beschreibung der zeitlichen Trajektorie der MEL-spektralen Merkmale des verhallten Sprachsignals statt. Das Beobachtungsmodell ist im Grunde ein Spezialfall des in Kap. 5.2.4 hergeleiteten rekursiven Modells und wird in [TN04] als Prädiktion erster Ordnung bezeichnet. Dabei wird die Auswirkung des Nachhalls auf den aktuell gültigen HMM-Zustand aufgrund des unmittelbar vorher beobachteten MEL-spektralen Merkmals des verhallten Sprachsignals berechnet.

Eine weitere Variante der dynamischen Adaption erhält man, indem die Adaption der akustischen Modelle nicht mit der Segmentrate, sondern deutlich seltener durchgeführt wird. So wird beispielsweise in [HF08] die Adaption auf den Nachhall unter anderem mit einer Adaption auf die Hintergrundstörung kombiniert. Die mittlere Leistung der Hintergrundstörung wird dabei zunächst mit einer *VAD* innerhalb von Sprachpausen geschätzt, so dass unmittelbar vor dem Einsetzen der Sprache die bereits auf den Nachhall angepassten akustischen Modelle zusätzlich auf die Hintergrundstörung adaptiert werden können.

3.3.2. Modifikation des Decoders

Der Effekt des Nachhalls kann schließlich auch erst bei der Dekodierung der Merkmalsvektorsequenz berücksichtigt werden. Dies geschieht in [SZK06] beispielsweise durch eine Modifikation des VITERBI-Algorithmus zur vereinfachten Dekodierung. Das Verfahren basiert auf der Kombination des akustischen Modells, welches mit sauberen Sprachsignalen trainiert wurde, mit einem Modell zur statistischen Beschreibung der RIA im Merkmalsraum. Die ursprüngliche Herleitung des Verfahrens beschränkt sich auf die Dekodierung mit MEL-spektralen Merkmalen. Die Änderung des VITERBI-Algorithmus besteht nun darin, dass dabei parallel sowohl nach der optimalen HMM-Zustandssequenz als auch nach der zugehörigen optimalen Sequenz der MEL-spektralen Merkmalsvektoren des sauberen Sprachsignals gesucht wird. Dazu wird versucht, die gemeinsame Likelihood der Sequenz der MEL-spektralen Merkmalsvektoren des sauberen Sprachsignals und der Repräsentation der RIA im MEL-spektralen Bereich unter der Nebenbedingung zu maximieren, dass deren Faltung die beobachtete Sequenz der MEL-spektralen Merkmalsvektoren des verhallten Sprachsignals ergibt. Neben der Tatsache, dass die Dekodierung recht aufwendig ist, werden die Emissionsverteilungsdichtefunktionen einzelner HMM-Zustände durch GAUSS-Verteilungsdichtefunktionen beschrieben. Dies ist an sich schon eine deutliche Einschränkung der Modellierungsmöglichkeit durch das HMM, da gewöhnlich *GMM*s an Stelle von GAUSS-Verteilungsdichtefunktionen verwendet werden. Zusätzlich ist die Verwendung von GAUSS-Verteilungsdichtefunktionen für den MEL-spektralen Bereich recht ungünstig, da die Merkmale beispielsweise nur nichtnegative Werte annehmen können. Obwohl das Verfahren in [SMK10] auf den log-MEL-spektralen Bereich erweitert werden konnte, blieb die Einschränkung auf die Verwendung von GAUSS-Verteilungsdichtefunktionen statt *GMM*s bestehen. Ein weiteres Problem, das sowohl im MEL- als im log-MEL-spektralen Bereich vorhanden ist, ist die vorhandene Korrelation zwischen einzelnen Komponenten der Merkmalsvektoren. Als Folge dessen müssen anders als bei der Verwendung von *MFCC*s vollbesetzte statt diagonalen Kovarianzmatrizen für die Emissionsverteilungsdichtefunktionen der HMM-Zustände zugrunde gelegt werden, was den Rechenaufwand bei der Dekodierung deutlich erhöht.

Eine andere Variante des Decoders besteht in der Ausnutzung von Unsichersinformationen

bezüglich der beobachteten Merkmale des verhallten Sprachsignals [PBB04]. Der Erkenner nutzt für die Erkennung im Wesentlichen diejenigen Merkmale, welche durch den Nachhall nur geringfügig verändert wurden.

Schließlich ist eine Kombination einer signalbasierten Enthallung mit einer entsprechenden Modifikationen des Decoders möglich [DNW09]. Die Idee besteht prinzipiell darin, dass der Decoder den nach der Enthallung verbleibenden zeitvarianten Reststörungen Rechnung trägt. Dies geschieht durch eine geeignete Anpassung der Varianzen zugehörig zu Emissionsverteilungsdichtefunktionen einzelner HMM-Zustände.

4. Wissenschaftliche Ziele

Während in der Literatur bereits zahlreiche Verfahren für die modellbasierte Entstörung akustischer Merkmale im Hinblick auf eine rauschrobuste Spracherkennung existieren, welche auf dem BAYES'schen Prinzip basieren, besteht das Ziel der Arbeit in der Entwicklung eines analogen Konzeptes zur gemeinsamen Kompensation des Nachhalls und der Hintergrundstörungen. Das Hauptaugenmerk liegt jedoch primär auf der Berücksichtigung des Nachhalls.

Merkmalsbasierte Verfahren besitzen grundsätzlich den Vorteil, dass sie vollkommen unabhängig von der Art des verwendeten Spracherkenners betrieben werden können und daher in der Praxis ein hohes Maß an Flexibilität bieten. Sie können im Wesentlichen direkt zwischen die Merkmalsextraktion und den Spracherkenner geschaltet werden, ohne jegliche Modifikationen am Spracherkenner vornehmen zu müssen. Insbesondere wird dabei eine meist aufwendige und komplizierte Adaption der Modellparameter des Spracherkenners auf veränderte Einsatzumgebungen vermieden.

Als Merkmale werden die *MFCC*s betrachtet, da sie aufgrund ihrer perzeptuell orientierten und relativ einfachen Berechnung in der Praxis eine weite Verbreitung gefunden haben. Obwohl sich das in dieser Dissertation vorgeschlagene Verfahren im Prinzip mit einigen Abwandlungen auch direkt im Cepstrum realisieren ließe, d.h. in dem Merkmalsbereich, der auch für die automatische Spracherkennung genutzt wird, wird hier aus zwei Gründen vorgeschlagen, dieses bereits eine Ebene vorher, d.h. im log-MEL-Spektrum, anzuwenden. Die Gründe bestehen zum einen darin, dass die LMSKs im Gegensatz zu den *MFCC*s einen annähernd homogenen Wertebereich besitzen, was aus numerischen Gründen vorteilhaft ist. Zum anderen werden die Einflüsse der Störung und des Nachhalls auf einzelne MEL-Frequenzbänder approximativ unabhängig sein, wohingegen im Cepstrum diese unabhängigen Einflüsse durch die Anwendung der *DCT* auf alle *MFCC*s verteilt werden.

Als Grundprinzip zur Merkmalsverbesserung dient die BAYES'sche Inferenz, die es erlaubt, Wissen beruhend auf zwei unterschiedlichen Informationsquellen in einer statistisch optimalen Art zu nutzen. Zu den Informationsquellen zählt einerseits das A-priori-Wissen über die Eigenschaften des sauberen Sprachsignals sowie des Störsignals im Merkmalsbereich. Zur Modellierung der Eigenschaften des sauberen Sprachsignals wird von interagierenden autoregressiven, linearen Prädiktionsmodellen (engl. ***S**witching **L**inear **D**ynamic **M**odel*s (*SLDM*s)) ausgegangen. Insbesondere werden auch Modelle höherer Ordnungen betrachtet, um Korrelation zwischen zeitlich weiter auseinander liegenden Sprachmerkmalsvektoren zu berücksichtigen. In diesem Bereich konzentriert sich die Arbeit auf die Herleitung und Untersuchung von Algorithmen zum iterativen Training und insbesondere einer sinnvollen Initialisierung der entsprechenden Modellparameter.

Die andere Informationsquelle besteht in dem sogenannten Beobachtungsmodell, welches die gemeinsamen Auswirkungen des Nachhalls und der Hintergrundstörungen auf die Form

der Merkmalsvektoren beschreibt und dessen Herleitung einen weiteren Schwerpunkt der Arbeit bildet. Dabei muss insbesondere die Tatsache berücksichtigt werden, dass dazu im Allgemeinen Wissen über die Ausbreitung akustischer Signale vom Sprecher zum Mikrofon, beispielsweise in Form einer RIA, zur Verfügung stehen muss. In der Regel hängt diese von der Beschaffenheit des Raumes ab, benötigt viele Parameter zu ihrer Darstellung und ist zudem zeitvariant. Ein weiterer wichtiger Aspekt in diesem Zusammenhang ist die Annahme eines sogenannten "blinden" Szenarios, bei dem die Einsatzumgebung des Spracherkenners sowie die Positionen des gewünschten Sprechers und des Mikrophons innerhalb der Umgebung unbekannt sind. Auf eine blinde Schätzung der gesamten detaillierten RIA beruhend auf dem eingehenden Mikrophonsignal wird hier allerdings verzichtet, da diese in der Regel höchst sensitiv und fehleranfällig ist. Statt dessen wird von einem stark vereinfachten Modell der RIA ausgegangen, das nur zwei Parameter besitzt: die Nachhallzeit sowie die Energie der RIA. Diese können deutlich robuster aus dem eingehenden Mikrophonsignal geschätzt werden. So beschäftigt sich die Arbeit sehr genau mit der Fragestellung, wie zu gegebenen RIA-Modellparametern ein adäquates Beobachtungsmodell berechnet werden kann. Dazu zählen unter anderem die Berechnung der modellbasierten Repräsentation der Raumimpulsantwort im Merkmalsraum und die Berechnung der statistischen Eigenschaften des Modellierungsfehlers.

4.1. Gliederung der Arbeit

Der Kern der Arbeit gliedert sich in zwei Hauptabschnitte.

In Kap. 5 erfolgt zunächst eine detaillierte theoretische Herleitung des BAYES'schen Verfahrens zur Merkmalsverbesserung. Dabei werden zunächst in Kap. 5.1 die verwendeten A-priori-Modelle zur statistischen Charakterisierung der zeitlichen Trajektorie der akustischen Merkmale des sauberen Sprachsignals sowie des Hintergrundstörsignals eingeführt. Anschließend wird ein sogenannter *EM*-Algorithmus zum iterativen Training von *SLDM*s beliebiger Ordnung hergeleitet sowie ein neuartiges Verfahren zur Initialisierung der *SLDM*s-Parameter vorgestellt. In Kap. 5.2 wird das Beobachtungsmodell zur Beschreibung des funktionellen Zusammenhanges zwischen den beobachteten Merkmalen des verhallten und gestörten Sprachsignals und den Merkmalen des sauberen Sprachsignals und des Hintergrundstörsignals hergeleitet. Dies geschieht anfangs unter der Annahme einer bekannten, zeitinvarianten RIA. Im Anschluss wird diese Voraussetzung jedoch fallen gelassen, wobei nun von einem stark vereinfachten statistischen Modell der RIA ausgegangen wird. Dieses erlaubt zudem die Formulierung eines zeitlich rekursiven Beobachtungsmodells, welches danach vorgestellt wird. Ein weiterer Aspekt, dem sich dieses Kapitel widmet, ist eine adäquate Modellierung des Beobachtungsfehlers. Schließlich werden in Kap. 5.3 unterschiedliche Verfahren zur approximativen Inferenz präsentiert, welche im Wesentlichen auf einem iterativen KALMAN-Filter sowie Modellkombinationsalgorithmen basieren.

Kapitel 6 befasst sich mit experimentellen Untersuchungen bezüglich der Leistungsfähigkeit des vorgestellten Verfahrens zur Merkmalsverbesserung. Diese werden mit Hilfe von zwei unterschiedlichen Sprachdatenbanken, mit einerseits kleinem und andererseits großem Vokabular, durchgeführt, welche in Kap. 6.1 ausführlich beschrieben werden. Als Kriterium für die Beurteilung der Leistungsfähigkeit wird in dieser Arbeit die nach der Spracherkennung endgültig erzielte Wortfehlerrate herangezogen. Nach einer Darstellung von Referenz-

ergebnissen, welche ohne die Verwendung jeglicher Merkmalsverbesserung erzielt wurden, und Ergebnissen einiger ausgewählter alternativer Verfahren in Kap. 6.2 und Kap. 6.3 werden in Kap. 6.4 die Resultate zu Voruntersuchungen bezüglich des Beobachtungsmodells aufgeführt, wobei die Schätzung der Parameter des Beobachtungsfehlers im Vordergrund steht. Kapitel 6.5 stellt die erzielten Ergebnisse zur Merkmalsenthallung vor, wobei insbesondere der Einfluss des A-priori-Sprachmodells sowie der des Beobachtungsmodells auf die Leistungsfähigkeit der Merkmalsverbesserung analysiert werden. Schließlich liefert Kap. 6.6 Ergebnisse zur gemeinsamen Merkmalsenthallung und -enstörung.

Die Arbeit wird mit einer Zusammenfassung und einem Ausblick in Kap. 7 abgeschlossen.

5. Konzept der modellbasierten BAYES'schen Merkmalsverbesserung

In diesem Kapitel wird eine modellbasierte Merkmalsverbesserung basierend auf BAYES'scher Inferenz vorgestellt. Eine Verbesserung auf der Merkmalsebene profitiert im Allgemeinen davon, dass sie sich auf nur denjenigen Anteil der Information beschränken kann, der auch tatsächlich für die Erkennung relevant ist. Natürlich können daraus auch Nachteile dadurch entstehen, dass eventuell zur Verbesserung benötigte Information nicht mehr zur Verfügung steht, wobei in der Regel dieser Aspekt eine untergeordnete Rolle spielt.

Das Ziel des hier vorgestellten Ansatzes besteht in der Bestimmung einer Folge $\hat{\mathbf{x}}_{1:M}^{(\mathrm{s})}$ von Schätzungen der LMSK-Vektoren des sauberen Sprachsignals

$$\hat{\mathbf{x}}_m^{(\mathrm{s})} := \left(\hat{x}_{m,0}^{(\mathrm{s})}, \ldots, \hat{x}_{m,Q-1}^{(\mathrm{s})}\right)^T \tag{5.1}$$

basierend auf der Beobachtung der Folge $\mathbf{y}_{1:M}^{(\mathrm{s})}$ der Merkmalsvektoren des verhallten und gestörten Sprachsignals. Insbesondere soll dieses Ziel durch einen Online-Algorithmus umgesetzt werden, was bedeutet, dass für die Schätzung des Merkmalsvektors $\hat{\mathbf{x}}_m^{(\mathrm{s})}$ nur alle vergangenen, der aktuelle und insbesondere keine (oder nur sehr wenige) zukünftige Merkmalsvektoren des verhallten und gestörten Sprachsignals verwendet werden dürfen.

Für die Schätzung wird zugrunde gelegt, dass es sich bei den nicht beobachtbaren Merkmalsvektorfolgen $\mathbf{x}_{1:M}^{(\mathrm{s})}$ und $\mathbf{n}_{1:M}^{(\mathrm{s})}$ sowie der beobachtbaren Merkmalsvektorfolge $\mathbf{y}_{1:M}^{(\mathrm{s})}$ um Realisierungen von vektorwertigen Zufallsprozessen $\breve{\mathbf{x}}_{1:M}^{(\mathrm{s})}$, $\breve{\mathbf{n}}_{1:M}^{(\mathrm{s})}$ sowie $\breve{\mathbf{y}}_{1:M}^{(\mathrm{s})}$ handelt. Aus statistischer Sicht kann das Schätzproblem als gelöst angesehen werden, sobald die A-posteriori-Verteilungsdichtefunktion $p\left(\mathbf{x}_m^{(\mathrm{s})} \middle| \mathbf{y}_{1:m}^{(\mathrm{s})}\right)$ bekannt ist. Diese erlaubt die Bestimmung von auf verschiedenen Kriterien basierenden Schätzwerten. So lässt sich beispielsweise zeigen, dass derjenige Schätzwert $\hat{\mathbf{x}}_m^{(\mathrm{s})}$ für $\mathbf{x}_m^{(\mathrm{s})}$, welcher den mittleren quadratischen Schätzfehler minimiert, durch den bedingten Erwartungswert

$$\boldsymbol{\mu}_{\breve{\mathbf{x}}_m^{(\mathrm{s})}\middle|\mathbf{y}_{1:m}^{(\mathrm{s})}} := \mathrm{E}\left[\breve{\mathbf{x}}_m^{(\mathrm{s})} \middle| \breve{\mathbf{y}}_{1:m}^{(\mathrm{s})} = \mathbf{y}_{1:m}^{(\mathrm{s})}\right] \tag{5.2}$$

gegeben ist. In der englischsprachigen Literatur wird ein solcher Schätzwert als ***M**inimum **M**ean **S**quared **E**rror* (*MMSE*) *estimate* bezeichnet. In dem besonderen Fall, dass die A-posteriori-Verteilungsdichtefunktion GAUSS-förmig ist, entspricht die zugehörige Kovarianzmatrix

$$\boldsymbol{\Sigma}_{\breve{\mathbf{x}}_m^{(\mathrm{s})}\middle|\mathbf{y}_{1:m}^{(\mathrm{s})}} := \mathrm{E}\left[\left(\breve{\mathbf{x}}_m^{(\mathrm{s})} - \boldsymbol{\mu}_{\breve{\mathbf{x}}_m^{(\mathrm{s})}\middle|\mathbf{y}_{1:m}^{(\mathrm{s})}}\right)\left(\breve{\mathbf{x}}_m^{(\mathrm{s})} - \boldsymbol{\mu}_{\breve{\mathbf{x}}_m^{(\mathrm{s})}\middle|\mathbf{y}_{1:m}^{(\mathrm{s})}}\right)^T \middle| \breve{\mathbf{y}}_{1:m}^{(\mathrm{s})} = \mathbf{y}_{1:m}^{(\mathrm{s})}\right] \tag{5.3}$$

der Schätzfehlerkovarianzmatrix und kann daher als Maß der verbliebenen Unsicherheit bezüglich der Schätzung angesehen werden. Das primäre Ziel bei dem hier vorgeschlagenen Verfahren zur Merkmalsverbesserung wird daher im Wesentlichen darin bestehen, Schätzwerte $\hat{\mathbf{x}}_m^{(\mathrm{s})}$ und $\hat{\boldsymbol{\Sigma}}_{\check{\mathbf{x}}_m^{(\mathrm{s})}}$ für die ersten beiden zentralen Momente $\boldsymbol{\mu}_{\check{\mathbf{x}}_m^{(\mathrm{s})}\left|\mathbf{y}_{1:m}^{(\mathrm{s})}\right.}$ und $\boldsymbol{\Sigma}_{\check{\mathbf{x}}_m^{(\mathrm{s})}\left|\mathbf{y}_{1:m}^{(\mathrm{s})}\right.}$ der A-posteriori-Verteilungsdichtefunktion zu bestimmen. Allerdings werden für die Schätzung zusätzlich einige wenige zukünftige Beobachtungen mit berücksichtigt, wie im Folgenden erläutert wird.

Ausgehend von diesen ersten Überlegungen wird zunächst der erweiterte Merkmalsvektor

$$\mathbf{z}_m^{(\mathrm{s})} := \left[\left(\boldsymbol{\chi}_m^{(\mathrm{s})}\right)^T, \left(\mathbf{n}_m^{(\mathrm{s})}\right)^T\right]^T \tag{5.4}$$

mit

$$\boldsymbol{\chi}_m^{(\mathrm{s})} := \left[\left(\mathbf{x}_m^{(\mathrm{s})}\right)^T, \ldots, \left(\mathbf{x}_{m-L_\mathrm{C}+1}^{(\mathrm{s})}\right)^T\right]^T \tag{5.5}$$

definiert, welcher sich aus einer Menge von $L_\mathrm{C} \in \mathbb{N}$ aufeinander folgenden Merkmalsvektoren des sauberen Sprachsignals $\boldsymbol{\chi}_m^{(\mathrm{s})}$ und einem Merkmalsvektor der Störung $\mathbf{n}_m^{(\mathrm{s})}$ zusammensetzt. Der Grund für genau diese Definition wird etwas später ersichtlich. Unter Verwendung von BAYES'scher Inferenz wird nun eine rekursive Formulierung für die A-posteriori-Verteilungsdichtefunktion $p\left(\mathbf{z}_m^{(\mathrm{s})}\middle|\mathbf{y}_{1:m}^{(\mathrm{s})}\right)$ bezüglich der Zeit, d.h. bezüglich des Segmentindexes m, vorgestellt. Dabei ist zu beachten, dass die benötigte Verteilungsdichtefunktion $p\left(\mathbf{x}_m^{(\mathrm{s})}\middle|\mathbf{y}_{1:m}^{(\mathrm{s})}\right)$ durch Marginalisierung aus $p\left(\mathbf{z}_m^{(\mathrm{s})}\middle|\mathbf{y}_{1:m}^{(\mathrm{s})}\right)$ hervorgeht.

Die Rekursion vollzieht sich in zwei Schritten. Im ersten Schritt, der sogenannten Prädiktion, wird ausgehend von der A-posteriori-Verteilungsdichtefunktion $p\left(\mathbf{z}_{m-1}^{(\mathrm{s})}\middle|\mathbf{y}_{1:m-1}^{(\mathrm{s})}\right)$ für den Segmentindex m die prädiktive Verteilungsdichtefunktion von $\check{\mathbf{z}}_m^{(\mathrm{s})}$ bedingt auf die vergangenen Beobachtungen $\mathbf{y}_{1:m-1}^{(\mathrm{s})}$ durch

$$p\left(\mathbf{z}_m^{(\mathrm{s})}\middle|\mathbf{y}_{1:m-1}^{(\mathrm{s})}\right) = \int_{\mathbb{R}^Q} p\left(\mathbf{z}_m^{(\mathrm{s})}\middle|\mathbf{z}_{m-1}^{(\mathrm{s})}, \mathbf{y}_{1:m-1}^{(\mathrm{s})}\right) p\left(\mathbf{z}_{m-1}^{(\mathrm{s})}\middle|\mathbf{y}_{1:m-1}^{(\mathrm{s})}\right) d\mathbf{z}_{m-1}^{(\mathrm{s})} \tag{5.6}$$

ausgedrückt. Im zweiten Schritt, der sogenannten Aktualisierung, wird dann die gesuchte A-posteriori-Verteilungsdichtefunktion $p\left(\mathbf{z}_m^{(\mathrm{s})}\middle|\mathbf{y}_{1:m}^{(\mathrm{s})}\right)$ für den Segmentindex m mit der BAYES'schen Regel gemäß

$$p\left(\mathbf{z}_m^{(\mathrm{s})}\middle|\mathbf{y}_{1:m}^{(\mathrm{s})}\right) = \frac{p\left(\mathbf{y}_m^{(\mathrm{s})}\middle|\mathbf{z}_m^{(\mathrm{s})}, \mathbf{y}_{1:m-1}^{(\mathrm{s})}\right) p\left(\mathbf{z}_m^{(\mathrm{s})}\middle|\mathbf{y}_{1:m-1}^{(\mathrm{s})}\right)}{\int_{\mathbb{R}^Q} p\left(\mathbf{y}_m^{(\mathrm{s})}\middle|\tilde{\mathbf{z}}_m^{(\mathrm{s})}, \mathbf{y}_{1:m-1}^{(\mathrm{s})}\right) p\left(\tilde{\mathbf{z}}_m^{(\mathrm{s})}\middle|\mathbf{y}_{1:m-1}^{(\mathrm{s})}\right) d\tilde{\mathbf{z}}_m^{(\mathrm{s})}} \tag{5.7}$$

$$\propto p\left(\mathbf{y}_m^{(\mathrm{s})}\middle|\mathbf{z}_m^{(\mathrm{s})}, \mathbf{y}_{1:m-1}^{(\mathrm{s})}\right) p\left(\mathbf{z}_m^{(\mathrm{s})}\middle|\mathbf{y}_{1:m-1}^{(\mathrm{s})}\right) \tag{5.8}$$

berechnet.

Die Durchführung des ersten Teilschrittes erfordert die Kenntnis der Verteilungsdichtefunktion $p\left(\mathbf{z}_m^{(\mathrm{s})} \middle| \mathbf{z}_{m-1}^{(\mathrm{s})}, \mathbf{y}_{1:m-1}^{(\mathrm{s})}\right)$, welche im Wesentlichen eine statistische Prädiktion der Dynamik der Sprache und der Störung liefert. Unter der Annahme, dass die Sprache und die Störung unabhängig voneinander sind, lässt sich diese Verteilungsdichtefunktion als Produkt

$$p\left(\mathbf{z}_m^{(\mathrm{s})} \middle| \mathbf{z}_{m-1}^{(\mathrm{s})}, \mathbf{y}_{1:m-1}^{(\mathrm{s})}\right) = p\left(\boldsymbol{\chi}_m^{(\mathrm{s})} \middle| \boldsymbol{\chi}_{m-1}^{(\mathrm{s})}, \mathbf{y}_{1:m-1}^{(\mathrm{s})}\right) \cdot p\left(\mathbf{n}_m^{(\mathrm{s})} \middle| \mathbf{n}_{m-1}^{(\mathrm{s})}, \mathbf{y}_{1:m-1}^{(\mathrm{s})}\right) \tag{5.9}$$

darstellen. Im nächsten Abschnitt wird gezeigt, wie sich die beiden auftretenden Verteilungsdichtefunktionen mittels $p\left(\mathbf{x}_m^{(\mathrm{s})} \middle| \mathbf{x}_{m-L_{\mathrm{AR}}:m-1}^{(\mathrm{s})}\right)$ und $p\left(\mathbf{n}_m^{(\mathrm{s})}\right)$ approximieren lassen, wobei $L_{\mathrm{AR}} \leq L_{\mathrm{C}}$ vorausgesetzt wird. Diese beiden Verteilungsdichtefunktionen bilden das sogenannte *A-priori-Modell*.

Für den zweiten Teilschritt der Rekursion wird gemäß (5.8) die Verteilungsdichtefunktion $p\left(\mathbf{y}_m^{(\mathrm{s})} \middle| \mathbf{z}_m^{(\mathrm{s})}, \mathbf{y}_{1:m-1}^{(\mathrm{s})}\right)$ benötigt, welche den Zusammenhang zwischen den L_{C} vergangenen Merkmalsvektoren $\mathbf{x}_{m-L_{\mathrm{C}}+1:m}^{(\mathrm{s})}$ des sauberen Sprachsignals, dem des Störsignals, $\mathbf{n}_m^{(\mathrm{s})}$, allen vergangenen Merkmalsvektoren $\mathbf{y}_{1:m-1}^{(\mathrm{s})}$ des verhallten und gestörten Sprachsignals und dem aktuellen Merkmalsvektor $\mathbf{y}_m^{(\mathrm{s})}$ des verhallten und gestörten Sprachsignals beschreibt.

Aufgrund des dispersiven Effektes des Nachhalls wird ein Zusammenhang zwischen der Merkmalsvektorfolge $\mathbf{x}_{m-L_{\mathrm{C}}+1:m}^{(\mathrm{s})}$ und $\mathbf{y}_m^{(\mathrm{s})}$ bestehen, woran auch die Motivation für die Wahl des zusammengesetzten Merkmalsvektors $\mathbf{z}_m^{(\mathrm{s})}$ erkennbar wird. Wird dabei der Wert von L_{C} größer als $\hat{L}_H$ gewählt, wobei $\hat{L}_H$ eine von der RIA zwischen Sprecher und Mikrophon abhängige und das zeitliche Ausmaß der Dispersion beschreibende geschätzte Größe ist, so kann die Bedingung von $\check{\mathbf{y}}_m^{(\mathrm{s})}$ auf $\mathbf{y}_{1:m-1}^{(\mathrm{s})}$ vernachlässigt werden, ohne dass dabei ein zu großer Fehler entsteht

$$p\left(\mathbf{y}_m^{(\mathrm{s})} \middle| \mathbf{z}_m^{(\mathrm{s})}, \mathbf{y}_{1:m-1}^{(\mathrm{s})}\right) \approx p\left(\mathbf{y}_m^{(\mathrm{s})} \middle| \mathbf{x}_{m-\hat{L}_H:m}^{(\mathrm{s})}, \mathbf{n}_m^{(\mathrm{s})}\right). \tag{5.10}$$

Diese Verteilungsdichtefunktion bildet das *Beobachtungsmodell*, welches die Beobachtung mit den zu schätzenden Größen verknüpft.

Das gesamte Konzept der modellbasierten BAYES'schen Merkmalsverbesserung wird in Abb. 5.1 veranschaulicht. Die Güte und Effizienz der Merkmalsverbesserung wird natürlich stark vom verwendeten *A-priori-Modell* und *Beobachtungsmodell* abhängen. Diese Modelle werden in den folgenden Kapiteln 5.1 und 5.2 sehr ausführlich beschrieben. An dieser Stelle soll nur vorausgreifend erwähnt werden, dass das Beobachtungsmodell natürlich in hohem Maße durch die RIA zwischen Sprecher und Mikrophon bedingt ist, welche im Allgemeinen sehr viele Parameter besitzt und als unbekannt angenommen wird. Diesem Problem wird hier mit der Einführung eines stark vereinfachten Modells der RIA begegnet, welches nur die zwei Parameter τ_h und σ_h besitzt. Diese werden aus dem verhallten und gestörten Sprachsignal $y(l)$ blind geschätzt.

5.1. A-priori-Modell

In diesem Abschnitt werden die für die Dynamik der Sprache und der Störung verwendeten *A-priori-Modelle* beschrieben. Im Sinne einer Anpassung der Modelle an die Charakteristik

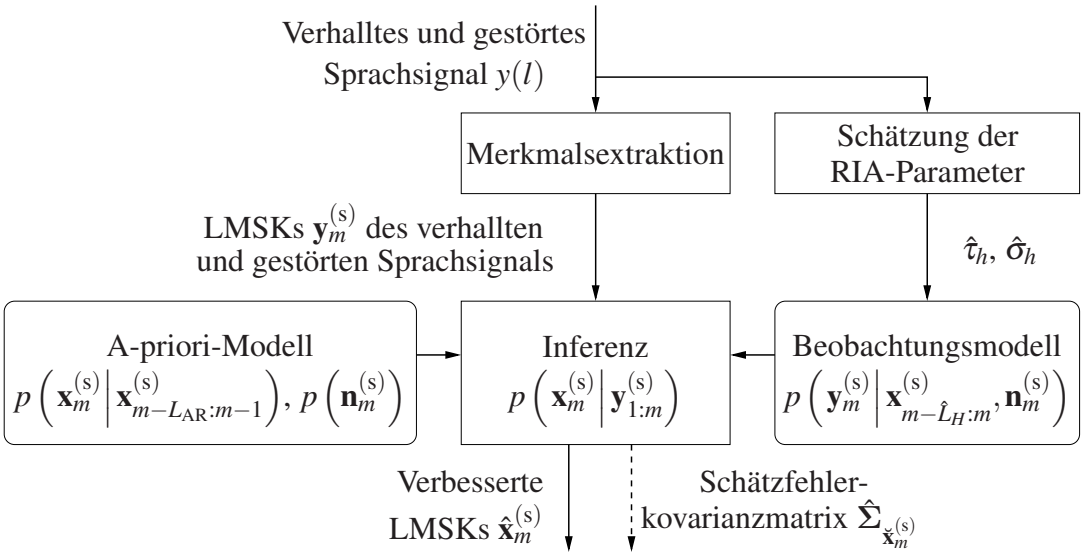


Abbildung 5.1.: *Blockschaltbild zur Veranschaulichung des Konzeptes der* BAYES*'schen Merkmalsverbesserung.*

des jeweiligen Signals und der Reduktion des Rechenaufwands durch eine niedrige Modellkomplexität werden unterschiedliche Arten von Modellen für die Sprache und die Störung vorgeschlagen.

5.1.1. Modell für die Sprache

Ein Sprachsignal ist in der Regel hochgradig instationär, denn die Änderungen im Signal entsprechen ja gerade der transportierten Information. Um das hohe Ausmaß der enthaltenen Dynamik explizit zu berücksichtigen, wird vorschlagen, die prädiktive Verteilungsdichtefunktion für die Merkmalsvektoren des sauberen Sprachsignals durch eine Mischung von $I \in \mathbb{N}$ unterschiedlichen, miteinander interagierenden Teilmodellen gemäß

$$p\left(\boldsymbol{\chi}_m^{(\mathrm{s})}\middle|\boldsymbol{\chi}_{m-1}^{(\mathrm{s})},\mathbf{y}_{1:m-1}^{(\mathrm{s})}\right) = \sum_{i=1}^{I} p\left(\boldsymbol{\chi}_m^{(\mathrm{s})}\middle|\boldsymbol{\chi}_{m-1}^{(\mathrm{s})},\mathbf{y}_{1:m-1}^{(\mathrm{s})},\zeta_m=i\right) P\left(\zeta_m=i\middle|\boldsymbol{\chi}_{m-1}^{(\mathrm{s})},\mathbf{y}_{1:m-1}^{(\mathrm{s})}\right) \tag{5.11}$$

zu modellieren. Dabei bezeichnet $\zeta_m \in \{1,\ldots,I\}$ eine Realisierung einer versteckten Zufallsvariablen $\breve{\zeta}_m$, deren Zustand das aktive Teilmodell zum Segmentindex m angibt. Bedingt durch die Definition des Merkmalsvektors $\boldsymbol{\chi}_m^{(\mathrm{s})}$ gemäß (5.5) lassen sich die teilmodellspezifischen Verteilungsdichtefunktionen $p\left(\boldsymbol{\chi}_m^{(\mathrm{s})}\middle|\boldsymbol{\chi}_{m-1}^{(\mathrm{s})},\mathbf{y}_{1:m-1}^{(\mathrm{s})},\zeta_m=i\right)$ vollständig nur unter Verwendung der Kenntnis der Verteilungsdichtefunktionen $p\left(\mathbf{x}_m^{(\mathrm{s})}\middle|\boldsymbol{\chi}_{m-1}^{(\mathrm{s})},\mathbf{y}_{1:m-1}^{(\mathrm{s})},\zeta_m=i\right)$ ausdrücken. Diese werden hier unter Vernachlässigung der Bedingung auf $\mathbf{y}_{1:m-1}^{(\mathrm{s})}$ durch lineare,

autoregressive Prädiktionsmodelle entsprechend

$$p\left(\mathbf{x}_m^{(\mathrm{s})}\middle|\boldsymbol{\chi}_{m-1}^{(\mathrm{s})},\mathbf{y}_{1:m-1}^{(\mathrm{s})},\zeta_m=i\right)\approx p\left(\mathbf{x}_m^{(\mathrm{s})}\middle|\mathbf{x}_{m-L_{\mathrm{AR}}:m-1}^{(\mathrm{s})},\zeta_m=i\right) \tag{5.12}$$

$$\approx\begin{cases}\mathcal{N}\left(\mathbf{x}_m^{(\mathrm{s})};\boldsymbol{\mu}_{\mathbf{x},i},\boldsymbol{\Sigma}_{\mathbf{x},i}\right) & \text{für } m\leq L_{\mathrm{AR}}\\ \mathcal{N}\left(\mathbf{x}_m^{(\mathrm{s})};\sum\limits_{\nu=1}^{L_{\mathrm{AR}}}\mathbf{A}_{i,\nu}\mathbf{x}_{m-\nu}^{(\mathrm{s})}+\mathbf{b}_i,\mathbf{V}_i\right) & \text{für } m>L_{\mathrm{AR}}.\end{cases} \tag{5.13}$$

approximiert. Gemäß dem i-ten Teilmodell gehen die Merkmalsvektoren $\mathbf{x}_m^{(\mathrm{s})}$ für Segmentindizes $m>L_{\mathrm{AR}}$ durch eine lineare Transformation aus ihren L_{AR} Vorgängern hervor, welche durch die Zustandsübergangsmatrizen $\mathbf{A}_{i,\nu}\in\mathbb{R}^{Q\times Q}$, $1\leq\nu\leq L_{\mathrm{AR}}$, und den Biaskorrekturvektor $\mathbf{b}_i\in\mathbb{R}^Q$ spezifiziert wird. Der verbleibende Prädiktionsfehler wird als Realisierung einer GAUSS-verteilten, mittelwertfreien Zufallsvariablen mit der Kovarianzmatrix $\mathbf{V}_i\in\mathbb{R}^{Q\times Q}$ betrachtet. Für Segmentindizes $m\leq L_{\mathrm{AR}}$ sind für eine derartige Prädiktion zu wenige Vorgänger vorhanden, so dass die Vorhersage mittels eines vergleichsweise einfachen *GMM*s mit den Mittelwertvektoren $\boldsymbol{\mu}_{\mathbf{x},i}\in\mathbb{R}^Q$ und den Kovarianzmatrizen $\boldsymbol{\Sigma}_{\mathbf{x},i}\in\mathbb{R}^{Q\times Q}$ erfolgt.

Für die Mischungsgewichte kann bei Vernachlässigung der Bedingung auf $\boldsymbol{\chi}_{m-1}^{(\mathrm{s})}$ unter der vereinfachten Annahme von zeitinvarianten Zustandsübergangswahrscheinlichkeiten

$$a_{k,i}:=P\left(\zeta_m=i\middle|\zeta_{m-1}=k\right)\qquad\text{für } m>L_{\mathrm{AR}} \tag{5.14}$$

die Approximation

$$P\left(\zeta_m=i\middle|\boldsymbol{\chi}_{m-1}^{(\mathrm{s})},\mathbf{y}_{1:m-1}^{(\mathrm{s})}\right)\approx P\left(\zeta_m=i\middle|\mathbf{y}_{1:m-1}^{(\mathrm{s})}\right) \tag{5.15}$$

$$\approx\begin{cases}\psi_i & \text{für } m\leq L_{\mathrm{AR}}\\ \sum\limits_{k=1}^{I}a_{k,i}P\left(\zeta_{m-1}=k\middle|\mathbf{y}_{1:m-1}^{(\mathrm{s})}\right) & \text{für } m>L_{\mathrm{AR}}\end{cases} \tag{5.16}$$

herangezogen werden, wobei

$$\psi_i:=P\left(\zeta_m=i\right)\qquad\text{für } m\leq L_{\mathrm{AR}} \tag{5.17}$$

die Zustandswahrscheinlichkeiten für die ersten L_{AR} Segmente angeben.

Ein derartiges Modell ist in der Literatur als schaltendes, lineares dynamisches Modell (engl. ***S**witching **L**inear **D**ynamic **M**odel* (*SLDM*)) [Kim94] bekannt. Es berücksichtigt explizit die zwischen aufeinanderfolgenden Merkmalsvektoren auftretenden Korrelationen, die einerseits durch die Spracherzeugung selbst bedingt sind und andererseits durch den Segmentüberlapp bei der Merkmalsextraktion entstehen. In welchem Maße die Korrelationen berücksichtigt werden, lässt sich durch die Ordnung L_{AR} des autoregressiven Modells steuern. Die Ordnung sollte natürlich von der Länge der Segmente zur Berechnung der Merkmalsvektoren abhängen. Für den hier betrachteten Fall der Merkmalsextraktion nach dem *ETSI*-Standard mit Parametern gemäß Tab. 2.1 sind Ordnungen der Größe 1 oder 2 typisch.

Die Parameter eines *SLDM* werden in der Regel unter Verwendung von Trainingsdatenbanken bestehend aus akustischen Äußerungen geschätzt. Dabei handelt es sich um sogenanntes unüberwachtes Modelltraining, da die Transkription des Sprachsignals bezüglich

der Zeitspannen der Aktivität einzelner Teilmodelle nicht vorhanden ist. In der Regel besteht sogar das Problem, dass die Anzahl der Teilmodelle sowie der Aspekt, welches Teilmodell überhaupt welche Dynamikbereiche modelliert, vollständig unbekannt ist. Auf das Training der *SLDM*s wird in Kap. 5.1.3 näher eingegangen.

5.1.2. Modell für die Störung

Die Charakteristik der Störung kann abhängig von der Umgebung stark variieren. Soll der Einsatzort des Spracherkenners möglichst uneingeschränkt sein, so müsste das Modell für die Störung alle möglichen Typen angemessen genau beschreiben können. Ein möglicher Lösungsweg, welcher jedoch eine sehr große und vielfältige Menge an Trainingsdaten erfordert, besteht darin, separate Modelle für jede einzelne Art der Störung aufzustellen. Das Kriterium zur Unterscheidung der Störungen könnte beispielsweise der Grad der Stationarität oder aber die entsprechende Frequenzcharakteristik sein. Während der Merkmalsverbesserung müsste dann basierend auf dem gestörten Signal das passende Modell gewählt werden.

Dieser Lösungsansatz wird hier jedoch aufgrund der hohen Anforderung auf die Menge und Vielfalt der Trainigsdaten nicht weiter verfolgt. Statt dessen wird hier von der vereinfachten Annahme ausgegangen, dass das Störsignal für kurze Zeitabschnitte, welche die Dauer einzelner Sprachäußerungen umfassen, seine Charakteristik nicht ändert. Diese Charakteristik ließe sich im Prinzip ebenfalls durch ein *SLDM* modellieren, wobei die entsprechenden Parameter durch die Verwendung einer *VAD* innerhalb von Sprachpausen geschätzt werden könnten. Obwohl zur Beschreibung der Störung in der Literatur bereits lineare dynamische Modelle eingesetzt wurden, wird hier aus zwei Gründen darauf verzichtet und statt dessen das Modell für die Störung (5.9) durch einen stationären weißen GAUSS'schen Zufallsprozess beschrieben:

$$p\left(\mathbf{n}_m^{(\mathrm{s})}\middle|\,\mathbf{n}_{m-1}^{(\mathrm{s})},\mathbf{y}_{1:m-1}^{(\mathrm{s})}\right)\approx p\left(\mathbf{n}_m^{(\mathrm{s})}\right)\approx\mathcal{N}\left(\mathbf{n}_m^{(\mathrm{s})};\boldsymbol{\mu}_{\mathbf{n}},\boldsymbol{\Sigma}_{\mathbf{n}}\right). \tag{5.18}$$

Der Mittelwertvektor $\boldsymbol{\mu}_{\mathbf{n}}$ und die Kovarianzmatrix $\boldsymbol{\Sigma}_{\mathbf{n}}$ werden dabei als konstant für die Dauer einer Sprachäußerung angenommen. Die Gründe für diese Wahl bestehen zum einen darin, dass *SLDM*s viele Modellparameter besitzen, so dass für eine zuverlässige Schätzung genügend lange Zeitabschnitte benötigt werden. Dieses verzögert das Nachführen der Modellparameter zwischen einzelnen Sprachäußerungen. Viel schwerwiegender ist zum anderen die Tatsache, dass durch die Verwendung eines *SLDM* die Stabilität der gesamten Merkmalsverbesserung gefährdet ist. Damit ist gemeint, dass es bei einem *SLDM* keine Beschränkung des Wertebereichs für den Schätzwert der Störung gibt, so dass bedingt durch das Zusammenspiel der rekursiven Art der Prädiktion durch ein *SLDM* und die auftretenden Schätzfehler die geschätzte Trajektorie der Störung vollkommen in die falsche Richtung verlaufen kann. Dieses Problem kann mit dem oben eingeführten Modell (5.18) nicht auftreten, da keine Korrelationen zwischen aufeinanderfolgenden Merkmalsvektoren der Störung angenommen werden und der Mittelwert $\boldsymbol{\mu}_{\mathbf{n}}$ über der Zeit konstant bleibt.

5.1.3. Training von *SLDM*s

Für die Bestimmung der *SLDM*-Parameter

$$\theta := \left\{ \boldsymbol{\mu}_{\mathbf{x},i}, \boldsymbol{\Sigma}_{\mathbf{x},i}, \mathbf{A}_{i,\nu}, \mathbf{b}_i, \mathbf{V}_i, \boldsymbol{\psi}_i, a_{i,k} \middle| i,k \in \{1,\ldots,I\}, \nu \in \{1,\ldots,L_{\mathrm{AR}}\} \right\} \tag{5.19}$$

wird von der gewöhnlich vorherrschenden Situation ausgegangen, dass die Trainingsdaten aus einer Menge von N unabhängigen Sprachäußerungen bestehen, welche durch die Menge der Merkmalsvektorsequenzen

$$\mathfrak{X} := \left\{ \mathbf{x}_{1:M_n}^{(n)} \middle| n \in \{1,\ldots,N\} \right\} \tag{5.20}$$

repräsentiert werden, wobei $\mathbf{x}_{1:M_n}^{(n)}$ die n-te Merkmalsvektorsequenz und M_n ihre Länge angibt. Dabei wird hier aus Gründen der Übersichtlichkeit auf die Kennzeichnung der Art der Merkmalsvektoren verzichtet.

Die bisher etablierte Methode zur Schätzung der Parametermenge θ besteht in der Anwendung des sogenannten *EM*-Algorithmus [DLR77]. Dabei handelt es sich um ein iteratives Verfahren zur lokalen Verbesserung einer initialen Parametermenge $\theta^{\{0\}}$, wobei das Kriterium in der Maximierung der sogenannten Likelihoodfunktion

$$\mathscr{L}(\theta) := p(\mathfrak{X}|\,\theta)\,. \tag{5.21}$$

besteht. Diese ist ein Maß für die Güte der Modellierung der Trainingsdaten mit Hilfe der Parametermenge θ und hängt daher insbesondere implizit von der Art des Modells zur Beschreibung der Dynamik in $\mathfrak{X}$ ab, was in dem hier betrachteten Fall das *SLDM* darstellt. Die direkte Auswertung der Likelihoodfunktion basierend auf einem *SLDM* würde die Kenntnis der zu der $\mathfrak{X}$ zugehörigen Menge

$$\mathfrak{Z} := \left\{ \boldsymbol{\zeta}_{1:M_n}^{(n)} \middle| n \in \{1,\ldots,N\} \right\} \tag{5.22}$$

von Zustandssequenzen erfordern, welche Auskunft über die Zeiträume der Aktivität einzelner Teilmodelle des *SLDM* geben. Da diese Zustandssequenzen nicht beobachtbar sind, wird statt der nicht realisierbaren, direkten Maximierung der Likelihoodfunktion $\mathscr{L}(\theta)$ ein Hilfsproblem betrachtet. Dazu wird im $(l+1)$-ten Iterationsschritt die Parametermenge

$$\theta^{\{l\}} := \left\{ \boldsymbol{\mu}_{\mathbf{x},i}^{\{l\}}, \boldsymbol{\Sigma}_{\mathbf{x},i}^{\{l\}}, \mathbf{A}_{i,\nu}^{\{l\}}, \mathbf{b}_i^{\{l\}}, \mathbf{V}_i^{\{l\}}, \boldsymbol{\psi}_i^{\{l\}}, a_{i,k}^{\{l\}} \middle| i,k \in \{1,\ldots,I\}, \nu \in \{1,\ldots,L_{\mathrm{AR}}\} \right\} \tag{5.23}$$

durch die Maximierung der Hilfsfunktion

$$\mathscr{Q}_{l+1}(\theta) := \mathrm{E}\left[\ln\left\{ p_{\check{\mathfrak{X}},\check{\mathfrak{Z}}}(\mathfrak{X},\mathfrak{Z}) \right\} \middle| \mathfrak{X}; \theta^{\{l\}} \right] \tag{5.24}$$

bestimmt. Das nicht vorhandene Wissen über die tatsächlichen Zustandssequenzen wird dabei durch eine weiche Entscheidung bezüglich der Aktivität einzelner Teilmodelle beruhend auf der alten Parametermenge $\theta^{\{l\}}$ approximiert. Die Anwendung des Logarithmus auf die Likelihoodfunktion dient der Vereinfachung der resultierenden Ausdrücke, wobei die Maximumstelle bedingt durch die strenge Monotonie des Logarithmus nicht verändert wird. Es kann gezeigt werden [DLR77], dass für die auf diese Weise mit dem *EM*-Algorithmus

für $l \geq 1$ iterativ bestimmten Parameterschätzungen $\theta^{\{l\}}$ die Likelihoodfunktion monoton wächst, d.h.

$$\mathscr{L}\left(\theta^{\{l\}}\right) \geq \mathscr{L}\left(\theta^{\{l-1\}}\right) \quad \forall l \in \mathbb{N}. \tag{5.25}$$

Insbesondere konvergiert $\theta^{\{l\}}$ für $l \to \infty$ mit der **W**ahr**s**cheinlich**k**eit (WSK) 1 gegen eine lokale Maximumstelle von $\mathscr{L}(\theta)$. Eine sehr ausführliche Herleitung für die Neuberechnung der Modellparameter für *SLDM*s beliebiger Ordnung gemäß dem *EM*-Algorithmus findet sich in Kap. A.1 im Anhang. An dieser Stelle werden der Vollständigkeit halber nur die resultierenden Formeln aufgeführt.

Zunächst werden die auf die Modellparameter $\theta^{\{l\}}$ bedingten Zustandswahrscheinlichkeiten

$$\eta_m^{(n,l)}(i) := P\left(\zeta_m^{(n)} = i \middle| \mathbf{x}_{1:M_n}^{(n)}; \theta^{\{l\}}\right) \tag{5.26}$$

$$\xi_m^{(n,l)}(k,i) := P\left(\zeta_m^{(n)} = i, \zeta_{m-1}^{(n)} = k \middle| \mathbf{x}_{1:M_n}^{(n)}; \theta^{\{l\}}\right) \tag{5.27}$$

geschickt durch eine abgewandelte Version des sogenannten BAUM-WELCH-Algorithmus berechnet (siehe Kap. A.1.1). Die zu $\theta^{\{l+1\}}$ gehörenden Parameter erhält man dann durch

$$\boldsymbol{\mu}_{\mathbf{x},i}^{\{l+1\}} = \frac{\sum_{n=1}^{N}\sum_{m=1}^{L_{\mathrm{AR}}} \eta_m^{(n,l)}(i)\,\mathbf{x}_m^{(n)}}{\sum_{n=1}^{N}\sum_{m=1}^{L_{\mathrm{AR}}} \eta_m^{(n,l)}(i)} \tag{5.28}$$

$$\boldsymbol{\Sigma}_{\mathbf{x},i}^{\{l+1\}} = \frac{\sum_{n=1}^{N}\sum_{m=1}^{L_{\mathrm{AR}}} \eta_m^{(n,l)}(i)\left(\mathbf{x}_m^{(n)} - \boldsymbol{\mu}_{\mathbf{x},i}^{\{l\}}\right)\left(\mathbf{x}_m^{(n)} - \boldsymbol{\mu}_{\mathbf{x},i}^{\{l\}}\right)^T}{\sum_{n=1}^{N}\sum_{m=1}^{L_{\mathrm{AR}}} \eta_m^{(n,l)}(i)} \tag{5.29}$$

$$\mathbf{V}_i^{\{l+1\}} = \frac{\sum_{n=1}^{N}\sum_{m=L_{\mathrm{AR}}+1}^{M_n} \eta_m^{(n,l)}(i)\left(\mathbf{x}_m^{(n)} - \sum_{\nu=1}^{L_{\mathrm{AR}}} \mathbf{A}_{i,\nu}^{\{l\}}\mathbf{x}_{m-\nu}^{(n)} - \mathbf{b}_i^{\{l\}}\right)\left(\mathbf{x}_m^{(n)} - \sum_{\nu=1}^{L_{\mathrm{AR}}} \mathbf{A}_{i,\nu}^{\{l\}}\mathbf{x}_{m-\nu}^{(n)} - \mathbf{b}_i^{\{l\}}\right)^T}{\sum_{n=1}^{N}\sum_{m=L_{\mathrm{AR}}+1}^{M_n} \eta_m^{(n,l)}(i)} \tag{5.30}$$

$$\psi_i^{\{l+1\}} = \frac{\sum_{n=1}^{N}\sum_{m=1}^{L_{\mathrm{AR}}} \eta_m^{(n,l)}(i)}{N \cdot L_{\mathrm{AR}}} \tag{5.31}$$

$$a_{k,i}^{\{l+1\}} = \frac{\sum_{n=1}^{N}\sum_{m=L_{\mathrm{AR}}+1}^{M_n} \xi_m^{(n,l)}(k,i)}{\sum_{n=1}^{N}\sum_{m=L_{\mathrm{AR}}+1}^{M_n} \eta_{m-1}^{(n,l)}(k)}. \tag{5.32}$$

Zur Berechnung der Zustandsübergangsmatrizen $\mathbf{A}_{i,\nu}^{\{l+1\}}$, $\nu \in \{1,\ldots,L_{\mathrm{AR}}\}$, und der Biaskor-

rekturvektoren $\mathbf{b}_i$ muss für jedes $i \in \{1,\ldots,I\}$ das lineare Gleichungssystem

$$\mathbf{G}_i^{\{l\}} \begin{bmatrix} \left(\mathbf{A}_{i,1}^{\{l+1\}}\right)^T \\ \vdots \\ \left(\mathbf{A}_{i,L_{\mathrm{AR}}}^{\{l+1\}}\right)^T \\ \left(\mathbf{b}_i\right)^T \end{bmatrix} = \mathbf{H}_i^{\{l\}}. \tag{5.33}$$

gelöst werden, wobei die darin auftretenden Matrizen $\mathbf{G}_i^{\{l\}} \in \mathbb{R}^{(L_{\mathrm{AR}}Q+1)\times(L_{\mathrm{AR}}Q+1)}$ und $\mathbf{H}_i^{\{l\}} \in \mathbb{R}^{(L_{\mathrm{AR}}Q+1)\times Q}$ gemäß

$$\mathbf{G}_i^{\{l\}} := \begin{bmatrix} \left\langle \mathbf{x}^{\{l\}}(i) \right\rangle_{L_{\mathrm{AR}}+1:M_n}^{[1,1]} & \cdots & \left\langle \mathbf{x}^{\{l\}}(i) \right\rangle_{L_{\mathrm{AR}}+1:M_n}^{[1,L_{\mathrm{AR}}]} & \left\langle \mathbf{x}^{\{l\}}(i) \right\rangle_{L_{\mathrm{AR}}+1:M_n}^{[1]} \\ \vdots & \ddots & \vdots & \vdots \\ \left\langle \mathbf{x}^{\{l\}}(i) \right\rangle_{L_{\mathrm{AR}}+1:M_n}^{[L_{\mathrm{AR}},1]} & \cdots & \left\langle \mathbf{x}^{\{l\}}(i) \right\rangle_{L_{\mathrm{AR}}+1:M_n}^{[L_{\mathrm{AR}},L_{\mathrm{AR}}]} & \left\langle \mathbf{x}^{\{l\}}(i) \right\rangle_{L_{\mathrm{AR}}+1:M_n}^{[L_{\mathrm{AR}}]} \\ \left(\left\langle \mathbf{x}^{\{l\}}(i) \right\rangle_{L_{\mathrm{AR}}+1:M_n}^{[1]}\right)^T & \cdots & \left(\left\langle \mathbf{x}^{\{l\}}(i) \right\rangle_{L_{\mathrm{AR}}+1:M_n}^{[L_{\mathrm{AR}}]}\right)^T & \sum_{n=1}^{N}\sum_{m=L_{\mathrm{AR}}+1}^{M_n} 1 \end{bmatrix} \tag{5.34}$$

$$\mathbf{H}_i^{\{l\}} := \begin{bmatrix} \left\langle \mathbf{x}^{\{l\}}(i) \right\rangle_{L_{\mathrm{AR}}+1:M_n}^{[1,0]} \\ \vdots \\ \left\langle \mathbf{x}^{\{l\}}(i) \right\rangle_{L_{\mathrm{AR}}+1:M_n}^{[L_{\mathrm{AR}},0]} \\ \left(\left\langle \mathbf{x}^{\{l\}}(i) \right\rangle_{L_{\mathrm{AR}}+1:M_n}^{[0]}\right)^T \end{bmatrix} \tag{5.35}$$

und die in den Matrizen auftretenden Elemente durch

$$\left\langle \mathbf{x}^{\{l\}}(i) \right\rangle_{m':m''}^{[\nu,o]} := \sum_{n=1}^{N}\sum_{m=m'}^{m''} \eta_m^{(n,l)}(i)\,\mathbf{x}_{m-\nu}^{(n)}\left(\mathbf{x}_{m-o}^{(n)}\right)^T \tag{5.36}$$

$$\left\langle \mathbf{x}^{\{l\}}(i) \right\rangle_{m':m''}^{[\nu]} := \sum_{n=1}^{N}\sum_{m=m'}^{m''} \eta_m^{(n,l)}(i)\,\mathbf{x}_{m-\nu}^{(n)} \tag{5.37}$$

definiert sind.

Gewöhnlich werden die Iterationen so lange ausgeführt, bis ein Abbruchkriterium erfüllt ist. Hier wird vorgeschlagen, die Iterationen abzubrechen, sobald die mittlere relative Verbesserung der Likelihoodfunktion pro einzelne Äußerung, welche durch

$$\delta_{\mathscr{L}}^{(l+1)} := \exp\left\{\frac{1}{N}\left[\ln\left[\mathscr{L}\left(\theta^{\{l+1\}}\right)\right] - \ln\left[\mathscr{L}\left(\theta^{\{l\}}\right)\right]\right]\right\} \tag{5.38}$$

definiert ist, eine vorgegebene Schranke $\varepsilon_{\mathscr{L}}$ unterschreitet. Dabei wird die mittlere Verbesserung $\delta_{\mathscr{L}}^{(l+1)}$ bewusst über Loglikelihoodfunktionen definiert, da die entsprechenden Likelihoodfunktionen so geringe Werte annehmen, dass sie numerisch nicht berechenbar sind.

Ein offensichtlicher Schwachpunkt des *EM*-Algorithmus besteht darin, dass er nur eine lokal optimale Lösung liefert. Zur Überwindung dieses Problems wurde in der bisher erschienenen Literatur unter anderem die sogenannte deterministische Abkühlung (engl. *deterministic annealing*) [UN98] vorgeschlagen, welche eine geeignete Modifikation des *EM*-Algorithmus vornimmt. Dabei wird eine Parallele zur statistischen Mechanik gezogen, die auf der Feststellung beruht, dass der Ausdruck für die negative Loglikelihoodfunktion $-\ln[\mathscr{L}(\theta)]$ äquivalent zu dem für die sogenannte freie Energie eines thermodynamischen Systems bei einer bestimmten festen Temperatur formuliert werden kann. In diesem Sinne kann die Maximierung der Likelihoodfunktion als Minimierung der freien Energie des entsprechenden Systems interpretiert werden. Das Besondere an der Feststellung dieser Analogie ist die Tatsache, dass sich die Minimierung in der Regel deutlich vereinfacht, wenn die Temperatur gegen den absoluten Nullpunkt strebt. Denn für den Grenzfall des absoluten Nullpunktes sind die Zustandswahrscheinlichkeiten (5.26) für alle $i \in \{1,\ldots,I\}$ gleich und hängen insbesondere nicht von $\theta^{\{I\}}$ ab. Daher besitzt dann die freie Energie als Funktion der Parametermenge θ nur eine einzige globale Minimumstelle, die mit der lokalen übereinstimmt und sofort angegeben werden kann. Durch die stetige Erhöhung der Temperatur findet eine stetige Deformation der Energiefunktion statt, bis sie beim Erreichen der Ausgangstemperatur in die negative Loglikelihoodfunktion übergeht, die gewöhnlich eine sehr komplexe Gestalt mit vielen lokalen Minimumstellen aufweist. Die Idee des Ansatzes liegt nun darin, für vom Nullpunkt bis zur Ausgangstemperatur wachsende, diskrete Temperaturen die lokalen Minimumstellen der Energiefunktion zu bestimmen und anzunehmen, dass man bedingt durch die stetige Deformation in jedem Schritt auch tatsächlich die globalen Minimumstellen erhält. Es sei jedoch betont, dass auch dieses Verfahren keine global optimale Lösung garantiert. Aufgrund dieses Problems sind Initialisierungsstrategien erforderlich, um eine geeignete Modellparametermenge $\theta^{\{0\}}$ zu bestimmen. Dieses Problem wird im nächsten Abschnitt behandelt.

Ein weiterer Nachteil des *EM*-Algorithmus besteht in dem verwendeten Kriterium der Maximierung der Likelihoodfunktion $\mathscr{L}(\theta)$ zur Berechnung der Parametermenge θ. Denn eine besonders gute Modellierung der Trainingsdaten durch ein *SLDM*, die in einem großen Wert der Likelihoodfunktion zum Ausdruck kommt, muss nicht zwangsweise zu einer besonders geringen Wortfehlerrate nach der Merkmalsverbesserung führen, die mit demselben *SLDM* durchgeführt wurde. Bedauerlicherweise existieren in der Literatur, soweit es dem Autor bekannt ist, bisher keine im Zusammenhang mit der Wortfehlerrate stehenden Kriterien zum Training von *SLDM*s. Ein möglicher Grund dafür liegt sicherlich in der sehr hohen Komplexität derartiger Kriterien bedingt durch die notwendige Berücksichtigung der Struktur des Erkenners sowie des gesamten Prozesses der Merkmalsverbesserung. Basierend auf diesen Ausführungen wird in dieser Arbeit trotz der angesprochenen Diskrepanz der *EM*-Algorithmus verwendet.

5.1.4. Initialisierung von *SLDM*-Parametern

Dem Thema der Initialisierung von *SLDM*-Parametern wurde in der Literatur bisher nur mäßige Beachtung geschenkt. Dabei sind die dafür soweit vorhandenen Methoden insofern unzufriedenstellend, als dass sie nicht speziell für die Initialisierung von *SLDM*-Parametern entwickelt wurden, sondern sich eher behelfsmäßig an Verfahren zur Initialisierung von *GMM*-Parametern orientieren. Soweit dem Autor bekannt existieren hauptsächlich zwei An-

sätze, die in unterschiedlichen Variationen ausgeführt werden können.

Beim ersten Ansatz wird die Anzahl der Teilmodelle von 1 bis zu der gewünschten Anzahl I iterativ erhöht. Die Methode beruht auf der Tatsache, dass im Falle nur eines Teilmodells sich diejenigen *SLDM*-Parameter θ, welche die Likelihoodfunktion $\mathscr{L}(\theta)$ maximieren, in einem Schritt direkt berechnen lassen. Denn bei nur einem vorhandenen Modell stellt sich offensichtlich die Frage nach den Zeiträumen der Modellaktivität nicht, so dass für die bedingte Modellwahrscheinlichkeit in (5.26) stets $\eta_m^{(n,l)}(i) = 1 \ \forall m,n,l$ gilt. Nun wird eine iterative Erhöhung der Teilmodellanzahl durch eine Spaltung der bisher gefundenen Teilmodelle vollzogen. Ein bestehendes Teilmodell gekennzeichnet durch die Modellparametermenge $\theta^{\{i\}}$ wird dabei jeweils in zwei neue Teilmodelle mit den Parametermengen $\theta^{\{i_1\}}$ und $\theta^{\{i_2\}}$ dadurch aufgeteilt, indem sowohl der *GMM*-Mittelwertvektor $\boldsymbol{\mu}_{\mathbf{x},i}$ als auch der Biaskorrekturvektor $\mathbf{b}_i$ jeweils in zwei entgegengesetzte Richtungen gemäß

$$\boldsymbol{\mu}_{\mathbf{x},i_1} := \boldsymbol{\mu}_{\mathbf{x},i} + \beta \mathbf{U}_{\boldsymbol{\Sigma}_{\mathbf{x},i}} \sqrt{\operatorname{diag}\left\{\boldsymbol{\Lambda}_{\boldsymbol{\Sigma}_{\mathbf{x},i}}\right\}} \tag{5.39}$$

$$\boldsymbol{\mu}_{\mathbf{x},i_2} := \boldsymbol{\mu}_{\mathbf{x},i} - \beta \mathbf{U}_{\boldsymbol{\Sigma}_{\mathbf{x},i}} \sqrt{\operatorname{diag}\left\{\boldsymbol{\Lambda}_{\boldsymbol{\Sigma}_{\mathbf{x},i}}\right\}} \tag{5.40}$$

und

$$\mathbf{b}_{i_1} := \mathbf{b}_i + \beta \mathbf{U}_{\mathbf{V}_i} \sqrt{\operatorname{diag}\left\{\boldsymbol{\Lambda}_{\mathbf{V}_i}\right\}} \tag{5.41}$$

$$\mathbf{b}_{i_2} := \mathbf{b}_i - \beta \mathbf{U}_{\mathbf{V}_i} \sqrt{\operatorname{diag}\left\{\boldsymbol{\Lambda}_{\mathbf{V}_i}\right\}} \tag{5.42}$$

mit einem Skalierungsfaktor $0 < \beta < 1$ perturbiert wird, wobei die Anwendung von $\operatorname{diag}\{\cdot\}$ auf eine Matrix derart zu verstehen ist, dass sie einen Vektor liefert, dessen Einträge aus den Elementen der Hauptdiagonalen der Matrix bestehen. Außerdem ist die Anwendung der Wurzel auf einen Vektor komponentenweise zu interpretieren. Die Richtungen der Verschiebungen werden durch Eigenwertzerlegungen der beiden Kovarianzmatrizen $\mathbf{V}_i$ und $\boldsymbol{\Sigma}_{\mathbf{x},i}$ entsprechend

$$\mathbf{V}_i = \mathbf{U}_{\mathbf{V}_i} \boldsymbol{\Lambda}_{\mathbf{V}_i} \mathbf{U}_{\mathbf{V}_i}^T \tag{5.43}$$

$$\boldsymbol{\Sigma}_{\mathbf{x},i} = \mathbf{U}_{\boldsymbol{\Sigma}_{\mathbf{x},i}} \boldsymbol{\Lambda}_{\boldsymbol{\Sigma}_{\mathbf{x},i}} \mathbf{U}_{\boldsymbol{\Sigma}_{\mathbf{x},i}}^T \tag{5.44}$$

bestimmt. Die Kovarianzmatrizen der neuen Teilmodelle werden beide gleich gemäß

$$\mathbf{V}_{i_1} = \mathbf{V}_{i_2} := \left(1 - \beta^2\right) \mathbf{V}_i \tag{5.45}$$

$$\boldsymbol{\Sigma}_{\mathbf{x},i_1} = \boldsymbol{\Sigma}_{\mathbf{x},i_2} := \left(1 - \beta^2\right) \boldsymbol{\Sigma}_{\mathbf{x},i} \tag{5.46}$$

herunter skaliert. Diese Wahl der Skalierung stellt sicher, dass die KULLBACK-LEIBLER-Divergenz zwischen den Verteilungsdichtefunktionen des Prädiktionsfehlers vor und nach der Modellspaltung minimiert wird. Bedingt durch die Erhöhung der Modellanzahl werden die Zustands- und Zustandsübergangswahrscheinlichkeiten ψ_i und $a_{i,k}$ derart angepasst, dass die durch die Spaltung entstandenen Teilmodelle jeweils die gleiche Wahrscheinlichkeit aufweisen. Die übrigen *SLDM*-Parameter bleiben bei der Spaltung unverändert.

In der Regel werden zwischen den einzelnen Spaltungen einige *EM*-Iteration zur Verfeinerung der neu entstandenen Teilmodelle durchgeführt. Variationen dieses Ansatzes unterscheiden sich weiterhin darin, ob in jedem Schritt alle vorhandenen oder nur die wahrscheinlichsten Teilmodelle gespaltet werden. Die beschriebene Art der iterativen Modellspaltung ist vom *GMM*-Training übernommen. Sie findet beispielsweise Einsatz im sogenannten

***H**idden* **M**ARKOV ***M**odell **T**oolkit* (*HTK*) [YEG+06], einer Programmbibliothek zur Erstellung und zum Training von HMMs, welche an der Universität Cambridge entwickelt wurde.

Der zweite Ansatz zur Initialisierung von *SLDM*s basiert auf der Idee einer initialen Clusterbildung [DBY07]. Zur Initialisierung von I *GMM*-Mittelwertvektoren $\boldsymbol{\mu}_{\mathbf{x},i}$ werden zunächst I Vektoren gemäß einer Gleichverteilung aus der Menge

$$\mathfrak{X}_{1:L_{\mathrm{AR}}} := \left\{ \mathbf{x}_m^{(n)} \middle| m \in \{1,\ldots,L_{\mathrm{AR}}\}, n \in \{1,\ldots,N\} \right\} \tag{5.47}$$

der L_{AR} ersten Merkmalsvektoren aller Trainingsäußerungen gezogen. Diese bilden die initialen Clusterzentren. Anschließend werden diese Zentren durch beispielsweise den sogenannten K-MEANS- oder FUZZY-K-MEANS-Algorithmus [DHS01] iterativ verbessert. Die Kovarianzmatrizen $\boldsymbol{\Sigma}_{\mathbf{x},i}$ sowie Zustandswahrscheinlichkeiten ψ_i lassen sich empirisch basierend auf einer harten Zuordnung der Elemente aus $\mathfrak{X}_{1:L_{\mathrm{AR}}}$ zu den Clusterzentren berechnen.

Eine Übertragung dieses Verfahrens auf *SLDM*s lässt sich dadurch bewerkstelligen, indem zunächts davon ausgegangen wird, dass das *SLDM* die Ordnung $L_{\mathrm{AR}} = 1$ besitzt und entsprechend alle Zustandsübergangsmatrizen $\mathbf{A}_{i,\nu}$ für $1 < \nu \leq L_{\mathrm{AR}}$ und $1 \leq i \leq I$ gleich der Nullmatrix gesetzt werden. Die Zustandsübergangsmatrix $\mathbf{A}_{i,1}$ wird zur Einheitsmatrix gesetzt und anschließend die initialen Biaskorrekturvektoren $\mathbf{b}_i$ durch die Gruppierung der Menge

$$\Delta_{\mathfrak{X}} := \left\{ \mathbf{x}_{m+1}^{(n)} - \mathbf{x}_m^{(n)} \middle| m \in \{1,\ldots,M_n - 1\}, n \in \{1,\ldots,N\} \right\} \tag{5.48}$$

bestehend aus den Differenzen aufeinanderfolgender Merkmalsvektoren bestimmt. Die Zustandsübergangswahrscheinlichkeiten $a_{i,k}$ sowie die Prädiktionsfehlerkovarianzmatrizen $\mathbf{V}_i$ werden auch hier empirisch durch eine harte Zuordnung der Vektoren aus der Menge $\Delta_{\mathfrak{X}}$ zu den einzelnen Teilmodellen ermittelt.

Der Nachteil der beiden Initialisierungsverfahren im Hinblick auf die Initialisierung der *SLDM*-Parameter besteht darin, dass sich alle berechneten Teilmodelle sehr stark ähneln, da insbesondere die Zustandsübergangsmatrizen $\mathbf{A}_{i,\nu}$ aller Teilmodelle gleich sind. Dieses widerspricht jedoch der Absicht, dass einzelne Teilmodelle möglichst unterschiedliche Dynamikbereiche der Sprachmerkmalsvektortrajektorie repräsentieren sollen.

Basierend auf dieser Diskrepanz wurde ein neuartiges stochastisches Verfahren zur Initialisierung von *SLDM*s entwickelt, welches bereits vom Autor in [KLHU+10] veröffentlicht wurde und in dieser Arbeit zum Teil erheblich modifiziert wurde. Es handelt sich dabei um ein stochastisches Verfahren, welches sehr stark an den K-MEANS++-Algorithmus [AV07] angelehnt ist und dessen Ziel darin besteht, möglichst signifikant unterschiedliche Teilmodelle zur Repräsentation der Trainingsdaten zu finden.

Genauer gesagt lässt sich die Initialisierung in zwei unabhängige Probleme aufteilen, wenn man von der nicht besonders einschränkenden Annahme ausgeht, dass die Zustandsübergangswahrscheinlichkeiten für den Segmentindex $m = L_{\mathrm{AR}}$ alle gleich sind, d.h.

$$P\left(\zeta_{L_{\mathrm{AR}}+1} = i \middle| \zeta_{L_{\mathrm{AR}}} = k\right) = \frac{1}{I} \;\; \forall i,k \in \{1,\ldots,I\}. \tag{5.49}$$

Während das erste Problem darin besteht, initiale *GMM*-Parameter $\boldsymbol{\mu}_{\mathbf{x},i}$, $\boldsymbol{\Sigma}_{\mathbf{x},i}$ und ψ_i zu finden, besteht das zweite Problem in der Bestimmung der initialen Parameter $\mathbf{A}_{i,\nu}$, $\mathbf{b}_i$, $\mathbf{V}_i$ und $a_{i,k}$ des autoregressiven dynamischen Modells. Im Folgenden werden Lösungsvorschläge für beide Probleme detailliert dargestellt.

Initialisierung der *GMM*-Parameter

Das hier vorgestellte Verfahren zur Initialisierung der *GMM*-Parameter ist durch Alg. 1 beschrieben und lässt sich in zwei Teile gliedern.

Im ersten Teil werden *GMM*-Mittelwertvektoren $\boldsymbol{\mu}_{\mathbf{x},i}$ gemäß der stochastischen Initialisierung des K-MEANS++Algorithmus bestimmt, wobei das Ziel in der Minimierung des Gesamtabstandes

$$D_{\text{INIT}} := \sum_{\mathbf{x}_m^{(n)} \in \mathfrak{X}_{1:L_{\text{AR}}}} \min_{1 \leq i \leq I} \left\| \boldsymbol{\mu}_{\mathbf{x},i} - \mathbf{x}_m^{(n)} \right\|^2 \tag{5.50}$$

besteht. Dabei werden nacheinander die *GMM*-Mittelwertvektoren $\boldsymbol{\mu}_{\mathbf{x},1}, \ldots, \boldsymbol{\mu}_{\mathbf{x},I}$ zufällig aus der Menge $\mathfrak{X}_{1:L_{\text{AR}}}$ gezogen. Das besondere an der K-MEANS++-artigen Initialisierung sind die Wahrscheinlichkeiten, die für die Ziehung einzelner Merkmalsvektoren verwendet werden. Bei der Ziehung des ersten *GMM*-Mittelwertvektors $\boldsymbol{\mu}_{\mathbf{x},1}$ sind alle Merkmalsvektoren aus $\mathfrak{X}_{1:L_{\text{AR}}}$ gleich wahrscheinlich, wobei dazu die Annahme verwendet wird, dass es keinen offensichtlichen Grund gibt, bestimmte Vektoren zu bevorzugen. Für alle weiteren Ziehungen werden die Wahrscheinlichkeiten für einzelne Merkmalsvektoren proportional zu ihrem minimalen quadratischen EUKLIDISCHEN Abstand zu allen bisher gezogenen *GMM*-Mittelwertvektoren gewählt. Durch diese Art der Wahl der Wahrscheinlichkeiten soll verhindert werden, dass Merkmalsvektoren, die zu nah an den bisher gezogenen *GMM*-Mittelwertvektoren liegen, als neue *GMM*-Mittelwertvektoren ausgewählt werden. Die stochastische Komponente des Algorithmus ist motiviert durch das Bestreben, die Wahrscheinlichkeit für die Wahl von eventuellen Ausreißern als *GMM*-Mittelwertvektoren zu minimieren, da Ausreißer per Definition natürlich einen großen Abstand zu allen Vektoren aufweisen, wobei ihre Anzahl jedoch sehr gering ist. Die Gesamtdistanz D_{INIT} kann für jede einzelne Initialisierung als Realisierung einer Zufallsvariablen $\tilde{D}_{\text{INIT}}$ angesehen werden, deren Erwartungswert das folgende Optimalitätskriterium [AV07]

$$\mathrm{E}\left[\tilde{D}_{\text{INIT}}\right] \leq 8\left[\ln(I)+2\right] D_{\text{INIT,OPT}} \tag{5.57}$$

erfüllt, wobei $D_{\text{INIT,OPT}}$ die minimal erreichbare Gesamtdistanz bei gegebener Menge der Merkmalsvektoren $\mathfrak{X}_{1:L_{\text{AR}}}$ bezeichnet.

Der zweite Teil des Algorithmus behandelt die Initialisierung der Kovarianzmatrizen $\boldsymbol{\Sigma}_{\mathbf{x},i}$ und Teilmo-dellwahrscheinlichkeiten ψ_i. Dazu erfolgt zunächst eine Zuordnung aller Merkmalsvektoren aus $\mathfrak{X}_{1:L_{\text{AR}}}$ zu den einzelnen *GMM*-Mittelwertvektoren $\boldsymbol{\mu}_{\mathbf{x},i}$. Mit den aus der Zuordnung resultierenden Clustern $\mathfrak{M}_i$ von Merkmalsvektoren lassen sich die Kovarianzmatrizen $\boldsymbol{\Sigma}_{\mathbf{x},i}$ als empirische Kovarianzmatrizen aller Vektoren in $\mathfrak{M}_i$ bezüglich $\boldsymbol{\mu}_{\mathbf{x},i}$ gemäß (5.55) und die Teilmodellwahrscheinlichkeiten ψ_i als relative Anzahl der Merkmalsvektoren in $\mathfrak{M}_i$ gemäß (5.56) berechnen.

Initialisierung der *SLDM*-Parameter

Für die Initialisierung der Parameter $\mathbf{A}_{i,\nu}$, $\mathbf{b}_i$, $\mathbf{V}_i$ und $a_{i,k}$ des autoregressiven dynamischen Modells wird die Ansatz der K-MEANS++-artigen Initialisierung geeignet modifiziert. Dabei wird jedoch die einschränkende Annahme gemacht, dass die Zustandsübergangsmatrizen $\mathbf{A}_{i,\nu}$ für $\nu > 1$ alle zur Nullmatrix gesetzt werden. Der Grund dafür wird an einer späteren

Algorithmus 1 Initialisierung der *GMM*-Parameter

Für $i = 1..I$

1. Ziehe einen Merkmalsvektor $\mathbf{x}_{m_i}^{(n_i)}$ aus der Menge $\mathfrak{X}_{1:L_{\mathrm{AR}}}$ zufällig mit der Wahrscheinlichkeit

$$P\left(\mathbf{x}_{m_i}^{(n_i)}\right) := \begin{cases} \frac{1}{N \cdot L_{\mathrm{AR}}} & \text{falls } i = 1 \\ \frac{D\left(\mathbf{x}_{m_i}^{(n_i)}\right)}{\sum_{n=1}^{N} \sum_{m=1}^{L_{\mathrm{AR}}} D\left(\mathbf{x}_m^{(n)}\right)} & \text{sonst} \end{cases}, \tag{5.51}$$

wobei

$$D\left(\mathbf{x}_m^{(n)}\right) := \min_{1 \le k \le i-1} \left\| \boldsymbol{\mu}_{\mathbf{x},k} - \mathbf{x}_m^{(n)} \right\|^2 \tag{5.52}$$

den minimalen quadratischen EUKLIDISCHEN Abstand des Merkmalsvektors $\mathbf{x}_m^{(n)}$ zu allen zuvor gezogenen *GMM*-Mittelwertvektoren bezeichnet.

2. Initialisiere den i-ten *GMM*-Mittelwertvektor durch $\boldsymbol{\mu}_{\mathbf{x},i} := \mathbf{x}_{m_i}^{(n_i)}$.

Ende für

Für $i = 1..I$

1. Berechne die Menge der zum i-ten Cluster zugeordneten Merkmalsvektoren

$$\mathfrak{M}_i := \left\{ \mathbf{x}_m^{(n)} \in \mathfrak{X}_{1:L_{\mathrm{AR}}} \,\middle|\, \Omega_m^{(n)} = i \right\} \tag{5.53}$$

mit

$$\Omega_m^{(n)} = \operatorname*{argmin}_{1 \le k \le I} \left\| \boldsymbol{\mu}_{\mathbf{x},k} - \mathbf{x}_m^{(n)} \right\|^2. \tag{5.54}$$

2. Initialisiere die Kovarianzmatrizen des Prädiktionsfehlers durch

$$\boldsymbol{\Sigma}_{\mathbf{x},i} = \frac{1}{|\mathfrak{M}_i|} \sum_{\mathbf{x}_m^{(n)} \in \mathfrak{M}_i} \left(\boldsymbol{\mu}_{\mathbf{x},i} - \mathbf{x}_m^{(n)} \right) \left(\boldsymbol{\mu}_{\mathbf{x},i} - \mathbf{x}_m^{(n)} \right)^T, \tag{5.55}$$

wobei $|\cdot|$ die Kardinalität einer Menge bezeichnet.

3. Initialisiere die Teilmodellwahrscheinlichkeiten durch

$$\psi_i := \frac{|\mathfrak{M}_i|}{N \cdot L_{\mathrm{AR}}}. \tag{5.56}$$

Ende für

Stelle ersichtlich. Dessen ungeachtet sind alle Zustandsübergangsmatrizen $\mathbf{A}_{i,1}$ nach der Initialisierung im Allgemeinen unterschiedlich.

Die Initialisierung lässt sich auch in diesem Fall in zwei Teile separieren.

Im ersten Teil, der in Alg. 2 dargestellt ist, werden nacheinander die Parametermengen $\mathfrak{S}_1, \ldots, \mathfrak{S}_I$ mit

$$\mathfrak{S}_i := \left\{ \mathbf{A}_{i,1}, \mathbf{b}_i, \mathbf{V}_i \right\} \tag{5.58}$$

auf Merkmalsvektorsequenzen $\mathbf{x}_{m_i:m_i+L_\mathrm{S}-1}^{(n_i)}$ der Länge L_S bestimmt, welche nacheinander zufällig aus der Menge aller möglichen Sequenzen

$$\mathfrak{X}_{\mathrm{SEQ},L_\mathrm{S}} := \left\{ \mathbf{x}_{m:m+L_\mathrm{S}-1}^{(n)} \middle| \, m \in \{L_\mathrm{AR}, \ldots, M_n - L_\mathrm{S} + 1\}, n \in \{1, \ldots, N\} \right\} \tag{5.59}$$

gezogen werden. Im Gegensatz zur *GMM*-Initialisierung, wo einzelne Merkmalsvektoren als Repräsentanten eines Clusters angesehen werden, werden hier nun die auf den gezogenen Merkmalsvektorsequenzen bestimmten Parametermengen $\mathfrak{S}_i$ als Repräsentanten eines Clusters verstanden.

Im Folgenden wird auf zwei Fragestellungen eingegangen, welche bei diesem Ansatz relevant sind:

1. Nach welchem Kriterium bestimmt man die Parametermenge $\mathfrak{S}_i$ beruhend auf der gewählten Sequenz $\mathbf{x}_{m_i:m_i+L_\mathrm{S}-1}^{(n_i)}$?

2. Nach welchem Kriterium wählt man die Sequenz $\mathbf{x}_{m_i:m_i+L_\mathrm{S}-1}^{(n_i)}$ zur Berechnung der Parameter $\mathfrak{S}_i$ aus?

Zur Lösung des ersten Problems wird hier vorgeschlagen, die Zustandsübergangsmatrix $\mathbf{A}_{i,1}$ und den Biaskorrekturvektor $\mathbf{b}_i$ mit Hilfe von linearer Regression auf der Merkmalsvektorsequenz $\mathbf{x}_{m_i:m_i+L_\mathrm{S}-1}^{(n_i)}$ zu bestimmen. Dazu wird die Lösung der kleinsten Quadrate des linearen Gleichungssystems (5.67) berechnet. Es sei ausdrücklich darauf hingewiesen, dass die Minimierung der Norm der Fehlerquadrate das Kriterium der Maximierung der Likelihoodfunktion $p\left(\mathbf{x}_{m_i:m_i+L_\mathrm{S}-1}^{(n_i)} \middle| \mathfrak{S}\right)$ impliziert, wenn man von der Nebenbedingung (5.68) an die Zustandsübergangsmatrix $\mathbf{A}_{i,1}$ absieht. Dieses lässt sich daran erkennen, dass sich das zur Maximierung der Likelihoodfunktion zu lösende, zuvor hergeleitete Gleichungssystem (5.33) unter Beachtung der Annahme $\mathbf{A}_{i,\nu} = \mathbf{0}$ für $\nu > 1$ und der Tatsache, dass nur ein Teilmodell für die Erzeugung der Sequenz $\mathbf{x}_{m_i:m_i+L_\mathrm{S}-1}^{(n_i)}$ verantwortlich ist, zu

$$\begin{bmatrix} \sum\limits_{m'=0}^{L_\mathrm{S}-2} \mathbf{x}_{m_i+m'}^{(n_i)} \left(\mathbf{x}_{m_i+m'}^{(n_i)}\right)^T & \sum\limits_{m'=0}^{L_\mathrm{S}-2} \mathbf{x}_{m_i+m'}^{(n_i)} \\ \sum\limits_{m'=0}^{L_\mathrm{S}-2} \left(\mathbf{x}_{m_i+m'}^{(n_i)}\right)^T & L_\mathrm{S}-1 \end{bmatrix} \begin{bmatrix} \mathbf{A}_{i,1}^T \\ \mathbf{b}_i^T \end{bmatrix} = \begin{bmatrix} \sum\limits_{m'=0}^{L_\mathrm{S}-2} \mathbf{x}_{m_i+m'}^{(n_i)} \left(\mathbf{x}_{m_i+m'+1}^{(n_i)}\right)^T \\ \sum\limits_{m'=0}^{L_\mathrm{S}-2} \left(\mathbf{x}_{m_i+m'+1}^{(n_i)}\right)^T \end{bmatrix} \tag{5.60}$$

reduziert. Dabei handelt es sich jedoch um die sogenannte Normalengleichung zum Gleichungssystem (5.67), weshalb die Lösung der kleinsten Quadrate von (5.67) implizit eine Lösung von (5.60) darstellt.

Algorithmus 2: Initialisierung der *SLDM*-Parameter (Teil 1)

Setze $i := 1$.

Solange $(i \leq I)$

1. Ziehe eine Merkmalsvektorsequenz $\mathbf{x}_{m_i:m_i+L_\mathrm{S}-1}^{(n_i)}$ der Länge L_S aus der Menge $\mathfrak{X}_{\mathrm{SEQ},L_\mathrm{S}}$ aller möglichen Merkmalsvektorsequenzen mit der Wahrscheinlichkeit

$$P\left(\mathbf{x}_{m_i:m_i+L_\mathrm{S}-1}^{(n_i)}\right) := \begin{cases} \dfrac{1}{\sum\limits_{n=1}^{N} M_n - L_\mathrm{S} - L_\mathrm{AR} + 2} & \text{falls } i = 1 \\ \dfrac{D_{\mathfrak{S}_{1:i-1}}\left(\mathbf{x}_{m_i:m_i+L_\mathrm{S}-1}^{(n_i)}\right)}{\sum\limits_{n=1}^{N} \sum\limits_{m=L_\mathrm{AR}}^{M_n-L_\mathrm{S}+1} D_{\mathfrak{S}_{1:i-1}}\left(\mathbf{x}_{m:m+L_\mathrm{S}-1}^{(n)}\right)} & \text{sonst} \end{cases}, \tag{5.61}$$

wobei

$$D_{\mathfrak{S}_{1:i-1}}\left(\mathbf{x}_{m:m+L_\mathrm{S}-1}^{(n)}\right) := \min_{1\leq k\leq i-1} D_{\mathfrak{S}_{k|1:i-1}}\left(\mathbf{x}_{m:m+L_\mathrm{S}-1}^{(n)}\right) \tag{5.62}$$

mit

$$D_{\mathfrak{S}_{k|1:i-1}}\left(\mathbf{x}_{m:m+L_\mathrm{S}-1}^{(n)}\right) := \max\left\{-\ln\left[\frac{p\left(\mathbf{x}_{m:m+L_\mathrm{S}-1}^{(n)}\middle|\,\mathfrak{S}_k\right)}{\max\limits_{1\leq i'\leq i-1} p\left(\mathbf{x}_{m_{i'}:m_{i'}+L_\mathrm{S}-1}^{(n_{i'})}\middle|\,\mathfrak{S}_{i'}\right)}\right], 0\right\} \tag{5.63}$$

$$p\left(\mathbf{x}_{m:m+L_\mathrm{S}-1}^{(n)}\middle|\,\mathfrak{S}_k\right) := \prod_{o=1}^{L_\mathrm{S}-1} \mathcal{N}\left(\mathbf{x}_{m+o}^{(n)}; \mathbf{A}_{k,1}\mathbf{x}_{m+o-1}^{(n)} + \mathbf{b}_k, \mathbf{V}_k\right) \tag{5.64}$$

$$= \prod_{o=1}^{L_\mathrm{S}-1} \mathcal{N}\left(\mathbf{e}_{m+o,k}^{(n)}; \mathbf{0}, \mathbf{V}_k\right) \tag{5.65}$$

$$\mathbf{e}_{m+o,k}^{(n)} := \mathbf{x}_{m+o}^{(n)} - \left(\mathbf{A}_{k,1}\mathbf{x}_{m+o-1}^{(n)} + \mathbf{b}_k\right) \tag{5.66}$$

den minimalen Abstand der Sequenz $\mathbf{x}_{m:m+L_\mathrm{S}-1}^{(n)}$ zu der Menge der bisher initialisierten Teilmodelle $\mathfrak{S}_{1:i-1}$ bezeichnet.

2. Berechne die Zustandsübergangsmatrix $\mathbf{A}_{i,1}$ und den Biaskorrekturvektor $\mathbf{b}_i$ als Lösung der kleinsten Quadrate des linearen Gleichungssystems

$$\begin{bmatrix}\mathbf{A}_{i,1} & \mathbf{b}_i\end{bmatrix}\begin{bmatrix}\mathbf{x}_{m_i}^{(n_i)} & \cdots & \mathbf{x}_{m_i+L_\mathrm{S}-2}^{(n_i)} \\ 1 & \cdots & 1\end{bmatrix} = \begin{bmatrix}\mathbf{x}_{m_i+1}^{(n_i)} & \cdots & \mathbf{x}_{m_i+L_\mathrm{S}-1}^{(n_i)}\end{bmatrix} \tag{5.67}$$

unter der Nebenbedingung

$$\mathbf{A}_{i,1}[r,s] = 0 \text{ für } |r-s| > \left\lfloor\frac{L_\mathrm{S}}{2}\right\rfloor - 2, \tag{5.68}$$

wobei $\lfloor\cdot\rfloor$ die Rundung auf die nächstkleinere oder gleich große, ganze Zahl bedeutet und $\mathbf{A}_{i,1}[r,s]$ das Element in der r-ten Zeile und s-ten Spalte der Matrix $\mathbf{A}_{i,1}$ bezeichnet.

3. Berechne die Kovarianzmatrix des Prädiktionsfehlers $\mathbf{V}_i$ gemäß

$$\mathbf{V}_i = \frac{1}{L_\mathrm{S}-1} \sum_{o=1}^{L_\mathrm{S}-1} \left[\left(\mathbf{e}_{m_i+o,i}^{(n_i)} \right) \left(\mathbf{e}_{m_i+o,i}^{(n_i)} \right)^T + \varepsilon_\mathbf{V} \cdot \mathrm{diag} \left\{ \left\| \mathbf{e}_{m_i+o,i}^{(n_i)} \right\|^2, \ldots, \left\| \mathbf{e}_{m_i+o,i}^{(n_i)} \right\|^2 \right\} \right]. \tag{5.69}$$

   mit einem Regularisierungsfaktor $0 < \varepsilon_\mathbf{V} \ll 1$.

4. **Für** $k = 1..i$

   a) Berechne die Menge der zum k-ten Teilmodell zugeordneten Merkmalsvektorsequenzen

$$\mathfrak{M}_{\mathrm{SEQ},k}(i) := \left\{ \mathbf{x}_{m:m+L_\mathrm{S}-1}^{(n)} \in \mathfrak{X}_{\mathrm{SEQ},L_\mathrm{S}} \,\middle|\, \Omega_{\mathrm{SEQ},m}^{(n)}(i) = k \right\} \tag{5.70}$$

   mit

$$\Omega_{\mathrm{SEQ},m}^{(n)}(i) := \underset{1 \le i' \le i}{\operatorname{argmin}} D_{\mathfrak{S}_{i'|1:i}} \left(\mathbf{x}_{m:m+L_\mathrm{S}-1}^{(n)} \right). \tag{5.71}$$

   b) Berechne die Teilmodellwahrscheinlichkeiten empirisch durch

$$P_k := \frac{\left| \mathfrak{M}_{\mathrm{SEQ},k}(i) \right|}{\sum_{n=1}^{N} M_n - L_\mathrm{S} - L_\mathrm{AR} + 2}. \tag{5.72}$$

   Ende für

5. Berechne die maximale Teilmodellwahrscheinlichkeit $P_{MAX} := \max_{1 \le k \le i} P_k$ und die Indexmenge aller wohl repräsentierten Teilmodelle

$$\mathscr{I} := \left\{ k | 1 \le k \le i, P_k \ge \varepsilon_{P,\mathrm{REL}} \cdot P_{MAX} \right\}. \tag{5.73}$$

   wobei $\varepsilon_{P,\mathrm{REL}}$ eine Konstante mit $0 < \varepsilon_{P,\mathrm{REL}} < 1$ bezeichnet.

6. Verwerfe alle Teilmodelle $k \notin \mathscr{I}$ und vergib neue, eindeutige Indizes $\{1, \ldots, |\mathscr{I}|\}$ an die Teilmodelle $k \in \mathscr{I}$.

7. Setze $i := |\mathscr{I}| + 1$.

Ende solange

Die Kovarianzmatrix $\mathbf{V}_i$ wird gemäß (5.69) berechnet, wobei der zweite Summand in (5.69) einen Regularisierungsterm darstellt, welcher die Invertierbarkeit von $\mathbf{V}_i$ gewährleistet. Sieht man von diesem Term ab, so verläuft die Berechnung der Kovarianzmatrix ebenfalls im Sinne der Maximierung der Likelihoodfunktion $p\left(\mathbf{x}_{m_i:m_i+L_\mathrm{S}-1}^{(n_i)} \middle| \mathfrak{S} \right)$, was aus einem Vergleich von (5.69) und (5.30) ersichtlich wird.

Das zweite Problem, nämlich die Wahl der Merkmalsvektorsequenz $\mathbf{x}_{m_i:m_i+L_\mathrm{S}-1}^{(n_i)}$, wird hier mit demselben stochastischen Prinzip angegangen, dass der K-MEANS++-artigen Initialisierung zugrunde liegt. Dazu kommen bei der Ziehung der ersten Sequenz $\mathbf{x}_{m_1:m_1+L_\mathrm{S}-1}^{(n_1)}$ alle Sequenzen in $\mathfrak{X}_{\mathrm{SEQ},L_\mathrm{S}}$ gleich wahrscheinlich in Betracht, während die Wahrscheinlichkeit für die Sequenzen zur Bestimmung der weiteren Parametermengen $\mathfrak{S}_i$, $i > 1$, proportional

zu ihrem minimalen Abstand (5.62) zu den bisher initialisierten Teilmodellen gesetzt wird. Der Abstand einer Merkmalsvektorsequenz $\mathbf{x}_{m:m+L_S-1}^{(n)}$ zu einem durch die Parametermenge $\mathfrak{S}_k$, $1 \leq k < i$ definierten, bereits bestimmten Teilmodell wird dabei durch die negative, normierte und nach unten durch Null beschränkte Loglikelihoodfunktion (5.63) definiert. Diese Wahl lässt sich anschaulich derart interpretieren, dass die negative Loglikelihoodfunktion $-\ln\left[p\left(\mathbf{x}_{m:m+L_S-1}^{(n)}\middle|\,\mathfrak{S}_k\right)\right]$ umso größere Werte annimmt, je schlechter die Sequenz $\mathbf{x}_{m:m+L_S-1}^{(n)}$ durch die Modellparametermenge $\mathfrak{S}_i$ dargestellt wird. Die Normierung gewährleistet die Tatsache, dass die zur Berechnung der bisherigen Parametermengen $\mathfrak{S}_k$, $1 \leq k < i$, jeweils verwendeten Merkmalsvektorsequenzen $\mathbf{x}_{m_k:m_k+L_S-1}^{(n_k)}$ einen nichtnegativen Abstand erhalten, wobei die am besten modellierte Merkmalsvektorsequenz den Abstand Null erhält. Da es rein theoretisch möglich wäre, dass eine beliebige Merkmalsvektorsequenz in der Menge $\mathfrak{X}_{\mathrm{SEQ},L_S}$ durch ein bereits bestimmtes Teilmodell $\mathfrak{S}_k$ besser repräsentiert wird als die Sequenz $\mathbf{x}_{m_k:m_k+L_S-1}^{(n_k)}$ selbst, so dass der resultierende Abstand ein negatives Vorzeichen erhielte, wird in diesem Fall der Abstand zu Null gesetzt. Eine solche Sequenz würde daher nicht für die Bestimmung weiterer Teilmodelle in Betracht gezogen werden, da sie bereits mit zufriedenstellender Genauigkeit durch die bestehenden Teilmodelle beschrieben wäre.

Ein weiterer Aspekt, dem besondere Beachtung geschenkt werden muss, ist die Wahl der Länge L_S der Merkmalsvektorsequenzen. Dabei müssen zwei gegensätzliche Argumente beachtet werden. Einerseits sollte die Sequenzlänge L_S besonders groß gewählt werden, um die Unterbestimmtheit des zur linearen Regression verwendeten Gleichungssystems (5.67) im Sinne der Bestimmung von aussagekräftigen Teilmodellen zu vermeiden. Zudem sollte berücksichtigt werden, dass eine gewisse Mindestlänge bereits aufgrund der Trägheit des menschlichen Vokaltraktes sinnvoll ist. Andererseits ist eine kürzere Sequenzlänge zu bevorzugen, da die Approximation eines instationären stochastischen Prozesses, als dessen Realisierungen die Merkmalsvektorsequenzen per Annahme angesehen werden, durch die Verwendung eines einzelnen linearen autoregressiven Modells nur lokal sinnvoll ist.

Hier wird der lokalen Charakterisierung durch einzelne lineare Teilmodelle eine höhere Priorität beigemessen, da dieses insbesondere in Übereinstimmung mit dem Ziel der Initialisierung von möglichst unterschiedlichen Teilmodellen steht. Die zur Vermeidung der Unterbestimmtheit des Gleichungssystems (5.67) gestellte Anforderung an eine große Sequenzlänge wird deshalb durch die Einführung der Nebenbedingung (5.68) abgeschwächt, gemäß derer alle Zustandsübergangsmatrizen $\mathbf{A}_{i,1}$ nach der Initialisierung eine Bandstruktur aufweisen müssen. Eine solche Nebenbedingung beschränkt die Anzahl der zu initialisierenden Parameter erheblich. Sie ist jedoch auch aus physikalischer Sicht sinnvoll, wenn, wie hier, beliebige Arten von *spektralen* Sprachmerkmalsvektoren betrachtet werden, bei denen Korrelationen vorwiegend zwischen benachbarten Vektorkomponenten auftreten.

Den gleichen Zweck wie die Einführung der Nebenbedingung (5.68) verfolgt auch das zu Beginn von Kap. 5.1.4 angesprochene Nullsetzen aller Zustandsübergangsmatrizen $\mathbf{A}_{i,\nu}$ für $\nu > 1$, das als zusätzliche Nebenbedingung angesehen werden kann. Ohne dieses Vorgehen müsste die minimale Sequenzlänge L_S entsprechend L_{AR} mal so groß sein, um genügend Bestimmungsgleichungen zur Initialisierung aller Zustandsübergangsmatrizen zu erhalten.

Bedingt durch die stochastische Natur des Algorithmus kann gelegentlich die Situation auftreten, dass Ausreißer innerhalb der Menge der gezogenen Merkmalsvektorsequenzen auftreten. Um diesem Problem zu begegnen, werden nach der Initialisierung des i-ten Teilmodells zunächst die empirischen Teilmodellwahrscheinlichkeiten P_k, $1 \leq k \leq i$, durch eine

harte Zuordnung aller Sequenzen in $\mathfrak{X}_{\mathrm{SEQ},L_\mathrm{S}}$ zu den bisher initialisierten Teilmodellen gemäß (5.72) berechnet. Danach werden alle unterrepräsentierten Teilmodelle k, welche

$$P_k \leq \varepsilon_{P,\mathrm{REL}} \cdot \max_{1\leq k'\leq i} P_{k'} \tag{5.74}$$

erfüllen, verworfen. Die Konstante $\varepsilon_{P,\mathrm{REL}}$ mit $0 < \varepsilon_{P,\mathrm{REL}} < 1$ gibt dabei an, wie zahlreich ein Teilmodell mindestens im Verhältnis zum am besten repräsentierten Teilmodell vertreten sein sollte, um nicht verworfen zu werden.

Nachdem alle Parametermengen $\mathfrak{S}_i$, $1 \leq i \leq I$, bestimmt sind, lassen sich die verbleibenden Zustandsübergangswahrscheinlichkeiten $a_{k,i}$ gemäß Alg. 3 im Wesentlichen mit Hilfe einer Zuordnung von Tupeln zweier aufeinanderfolgender Sequenzen zu jeweils zwei aufeinanderfolgenden Teilmodellen berechnen. Man beachte hierbei die große Ähnlichkeit zur Verfeinerung der Zustandsübergangswahrscheinlichkeiten gemäß (5.32) bei der Durchführung des *EM*-Algorithmus.

Algorithmus 3 Initialisierung der *SLDM*-Parameter (Teil 2)

Für $i = 1..I$

- Berechne die Menge der zum i-ten Cluster zugeordneten Merkmalsvektorsequenzen $\mathfrak{M}_{\mathrm{SEQ},i}(I)$.

Ende für
Für $i = 1..I$

Für $k = 1..I$

a) Berechne die Menge von Merkmalsvektorsequenztupeln

$$\mathfrak{M}_{\mathrm{SEQ},k,i}(I) := \Big\{ \Big(\mathbf{x}^{(n)}_{m:m+L_\mathrm{S}-1}, \mathbf{x}^{(n)}_{m+1:m+L_\mathrm{S}}\Big) \Big| \, \Omega^{(n)}_{\mathrm{SEQ},m}(I) = k, \Omega^{(n)}_{\mathrm{SEQ},m+1}(I) = i, \\ m \in \{L_\mathrm{AR}, \ldots, M_n - L_\mathrm{S}\}, n \in \{1, \ldots, N\} \Big\}, \tag{5.75}$$

b) Setze die Zustandsübergangswahrscheinlichkeiten zu $a_{k,i} = \dfrac{\left|\mathfrak{M}_{\mathrm{SEQ},k,i}(I)\right|}{\left|\mathfrak{M}_{\mathrm{SEQ},k}(I)\right|}$.

Ende für

Ende für

5.2. Beobachtungsmodell

Im Folgenden wird ein Beobachtungsmodell hergeleitet, welches einen Zusammenhang zwischen den LMSK-Vektoren des verhallten und gestörten Sprachsignals, $\mathbf{y}^{(\mathrm{s})}_m$, und den LMSK-Vektoren des sauberen Sprachsignals, $\mathbf{x}^{(\mathrm{s})}_m$, sowie denjenigen des Störsignals, $\mathbf{n}^{(\mathrm{s})}_m$, beschreibt.

Den Ausgangspunkt dazu bildet die Beschreibung des verhallten und gestörten Sprachsignals $y(l)$ im Zeitbereich, wonach es durch die Überlagerung des verhallten Sprachsignals $s(l)$ mit einem Störsignal $n(l)$ entsteht, d. h.

$$y(l) = s(l) + n(l). \tag{5.76}$$

Das verhallte Sprachsignal $s(l)$ geht dabei aus der Faltung des sauberen Sprachsignals $x(l)$ mit der RIA $h(l)$ vom Sprecher zum Mikrophon hervor

$$s(l) = (x * h)(l) = \sum_{p'=0}^{L_h-1} h(p')x(l-p'). \tag{5.77}$$

Um zu einem handhabbaren analytischen Ausdruck im Beobachtungsmodell zu gelangen, wird dabei zunächst in (5.77) die vereinfachende Annahme gemacht, dass die RIA $h(l)$ zeitinvariant und kausal ist und mit einer endlichen Anzahl von Abtastwerten L_h ausreichend approximiert werden kann, d.h.

$$h(l) = 0 \quad \text{für} \quad l < 0 \ \wedge \ l \geq L_h. \tag{5.78}$$

Ausgehend von den Modellen (5.76) und (5.77) wird nun zunächst ein exakter Zusammenhang zwischen den Kurzzeit-Spektren der auftretenden Zeitsignale $x(l)$ und $n(l)$ und $y(l)$ dargelegt. Anschließend wird ein auf (5.76) basierender, approximativer Zusammenhang zwischen den LMSK-Vektoren $\mathbf{y}_m^{(s)}$, $\mathbf{x}_m^{(s)}$ und $\mathbf{n}_m^{(s)}$ formuliert, welcher die Grundlage für das Beobachtungsmodell bildet.

5.2.1. Zusammenhang im Zeit-Frequenz-Bereich

Aufgrund der Linearität der *DTSTFT* folgt aus (5.76) direkt

$$Y(m,k) = S(m,k) + N(m,k). \tag{5.79}$$

Um das Kurzzeit-Spektrum des verhallten Signals $S(m,k)$ durch das des sauberen Sprachsignals $X(m,k)$ darstellen zu können, muss die Annahme getroffen werden, dass die Kurzzeit-Spektren durch Überabtastung berechnet werden. Dieses bedeutet, dass die Parameter zur Berechnung der Kurzzeit-Spektren, nämlich die Länge des Analysefensters L_w, der Fenstervorschub B und die Anzahl der Frequenzbins K bei der *DFT*, so gewählt werden, dass sie die beiden Bedingungen

$$B \leq K \tag{5.80}$$

$$B \leq L_w \tag{5.81}$$

erfüllen. Weiterhin soll sogar von der stärkeren Bedingung

$$B \leq L_w \leq K \tag{5.82}$$

ausgegangen werden, die gemäß Tab. 2.1 bei der Merkmalsextraktion gemäß [ETSb] erfüllt wird.

Im Fall der Überabtastung lässt sich das verhallte Signal $s(l)$ mit Hilfe seiner GABOR-Reihe [WR90, FR94]

$$s(l) = \sum_{m=-\infty}^{\infty} \sum_{k=0}^{K-1} \sigma_{m,k} \cdot w_{\mathrm{S}}(l - mB) e^{j\frac{2\pi}{K}kl} \tag{5.83}$$

darstellen, wobei die GABOR-Koeffizienten $\sigma_{m,k}$ durch

$$\sigma_{m,k} = S(m,k) e^{-j\frac{2\pi}{K}kmB} \tag{5.84}$$

mit dem Kurzzeit-Spektrum zusammenhängen. Dabei bezeichnet $w_{\mathrm{S}}(l)$ ein zum Analysefenster biorthogonales Synthesefenster, welches die sogenannte Vollständigkeitsbedingung [FR94, Gl. (21)]

$$\sum_{m=-\infty}^{\infty} \sum_{k=0}^{K-1} w_{\mathrm{S}}(l - mB) w_{\mathrm{A}}(p' - mB) e^{j\frac{2\pi}{K}k(l-p')} = \delta\left(l - p'\right) \quad \text{für} \quad l, p' \in \mathbb{Z} \tag{5.85}$$

erfüllt. Unter der Annahme, dass das Synthesefenster den gleichen Träger wie das Analysefenster besitzt, d.h.

$$w_{\mathrm{S}}(l') = 0 \quad \text{für} \quad l' < 0 \ \wedge \ l' \geq L_w, \tag{5.86}$$

lässt sich zeigen, dass sich die Vollständigkeitsbedingung (5.85) zu

$$\sum_{m=-\infty}^{\infty} w_{\mathrm{S}}(l - mB) w_{\mathrm{A}}(l - mB) = \frac{1}{K} \quad \text{für} \quad 0 \leq l < B \tag{5.87}$$

vereinfacht. Der entsprechende Beweis sowie die Herleitung einer einfachen Vorschrift für die Berechnung eines Synthesefensters $w_{\mathrm{S}}(l')$ zu einem gegebenem Analysefenster $w_{\mathrm{A}}(l')$ findet sich in Kap. A.2.1 im Anhang. Da das Synthesefenster im Allgemeinen nicht eindeutig bestimmt ist, wird die Berechnung desjenigen Synthesefensters betrachtet, welches die kleinste ℓ^2-Norm besitzt und damit die größtmögliche Konzentration im Zeitbereich aufweist [QC93].

Wird (5.84) in (5.83) eingesetzt, ergibt sich

$$s(l) = \sum_{m=-\infty}^{\infty} w_{\mathrm{S}}(l - mB) \sum_{k=0}^{K-1} S(m,k) \cdot e^{j\frac{2\pi}{K}k(l-mB)} \tag{5.88}$$

und es lässt sich erkennen, dass sich das verhallte Signal $s(l)$ perfekt aus seinem Kurzzeit-Spektrum $S(m,k)$ rekonstruieren lässt. Diese Art der Rekonstruktion ist in der englischsprachigen Literatur unter dem Namen ***W**eighted **O**ver**l**ap **A**dd* (*WOLA*) bekannt [CR83].

Die bisher dargestellten Ergebnisse und insbesondere (5.88) gelten natürlich in gleicher Weise für das saubere Sprachsignal $x(l)$

$$x(l) = \sum_{m=-\infty}^{\infty} w_{\mathrm{S}}(l - mB) \sum_{k=0}^{K-1} X(m,k) \cdot e^{j\frac{2\pi}{K}k(l-mB)}. \tag{5.89}$$

Um nun zu einem Ausdruck von $S(m,k)$ in Abhängigkeit von $X(m,k)$ zu gelangen, wird zunächst $S(m,k)$ analog zu (2.3) unter Verwendung von (2.2) gemäß

$$S(m,k) = \sum_{l'=0}^{L_w-1} w_{\mathrm{A}}(l')s(l'+mB)\cdot e^{-j\frac{2\pi}{K}kl'} \tag{5.90}$$

dargestellt. Anschließend werden, den Ausführungen in [AC07b] folgend, nacheinander die Gleichungen (5.77) und (5.89) in (5.90) eingesetzt:

$$\begin{aligned} S(m,k) &= \sum_{l'=0}^{L_w-1} w_{\mathrm{A}}(l') \sum_{p'=0}^{L_h-1} h(p')x(l'+mB-p')\cdot e^{-j\frac{2\pi}{K}kl'} && (5.91)\\ &= \sum_{l'=0}^{L_w-1} w_{\mathrm{A}}(l') \sum_{p'=0}^{L_h-1} h(p') \left[\sum_{m'=-\infty}^{\infty} w_{\mathrm{S}}(l'+\left(m-m'\right)B-p') \right. \\ &\qquad \left. \cdot \sum_{k'=0}^{K-1} X(m',k')\cdot e^{j\frac{2\pi}{K}k'[l'+(m-m')B-p']} \right] \cdot e^{-j\frac{2\pi}{K}kl'}. && (5.92) \end{aligned}$$

Definiert man nun die Funktionen

$$h_{k,k'}(m'') := \sum_{p'=0}^{L_h-1} h(p') \sum_{l'=0}^{L_w-1} w_{\mathrm{A}}(l')w_{\mathrm{S}}(l'+m''B-p')\cdot e^{j\frac{2\pi}{K}k'[l'+m''B-p']}\cdot e^{-j\frac{2\pi}{K}kl'}, \tag{5.93}$$

welche im Folgenden für $k \neq k'$ als Kreuzbandfilter und für $k = k'$ als Band-zu-Band-Filter bezeichnet werden, so lässt sich (5.92) durch

$$\begin{aligned} S(m,k) &= \sum_{m'=-\infty}^{\infty} \sum_{k'=0}^{K-1} X(m',k')h_{k,k'}(m-m') && (5.94)\\ &= \sum_{k'=0}^{K-1} \sum_{m''=-\infty}^{\infty} X(m-m'',k')h_{k,k'}(m'') && (5.95) \end{aligned}$$

ausdrücken. Man erkennt an (5.95), dass zur Berechnung des Kurzzeit-Spektrums des verhallten Signals $S(m,k)$ zunächst in jedem Frequenzbin k' separat eine Faltung des Kurzzeit-Spektrums des unverhallten Signals $X(m,k')$ mit $h_{k,k'}(m)$ bezüglich m durchgeführt wird und anschließend alle Ergebnisse aufsummiert werden.

Der Betrag und damit der Einfluss der Kreuzbandfilter $h_{k,k'}(m)$ verringert sich mit wachsendem Abstand $|k-k'|$ mod K. Um dieses zu erkennen, wird zunächst die Funktion

$$\phi_{k,k'}(l) := \sum_{l'=0}^{L_w-1} w_{\mathrm{A}}(l')w_{\mathrm{S}}(l'+l)\cdot e^{j\frac{2\pi}{K}k'(l'+l)}\cdot e^{-j\frac{2\pi}{K}kl'} \tag{5.96}$$

definiert. Damit lässt sich (5.93) derart interpretieren, dass eine Funktion durch die Faltung zwischen der Impulsantwort $h(l)$ und $\phi_{k,k'}(l)$ gebildet wird

$$\tilde{h}_{k,k'}(l) := \left(h * \phi_{k,k'}\right)(l) \tag{5.97}$$

welche anschließend mit der Rate $\frac{1}{B}$ abgetastet wird

$$h_{k,k'}(m) = \tilde{h}_{k,k'}(mB) \tag{5.98}$$

$$= \sum_{p'=0}^{L_h-1} h(p')\phi_{k,k'}(l-p')\big|_{l=mB}. \tag{5.99}$$

Bildet man nun die zeitdiskrete FOURIER-Transformierte (engl. *DTFT*) von (5.97)

$$\tilde{H}_{k,k'}\left(e^{j\theta}\right) := \sum_{l=-\infty}^{\infty} \tilde{h}_{k,k'}(l)e^{-jl\theta}, \tag{5.100}$$

so folgt aus (5.97) mit dem Faltungssatz für die *DTFT* sofort

$$\tilde{H}_{k,k'}\left(e^{j\theta}\right) = H\left(e^{j\theta}\right)\Phi_{k,k'}\left(e^{j\theta}\right). \tag{5.101}$$

Wenn man in einem nächsten Schritt die Funktion $\phi_{k,k'}(l)$ als Faltung

$$\phi_{k,k'}(l) = \left(w_{\mathrm{MA},k} * w_{\mathrm{MS},k'}\right)(l) \tag{5.102}$$

von zwei modulierten Fensterfunktionen

$$w_{\mathrm{MA},k}(l) := w_{\mathrm{A}}(-l)\cdot e^{j\frac{2\pi}{K}kl} \tag{5.103}$$

$$w_{\mathrm{MS},k'}(l) := w_{\mathrm{S}}(l)\cdot e^{j\frac{2\pi}{K}k'l} \tag{5.104}$$

beschreibt, so folgt für die *DTFT* von $\phi_{k,k'}(l)$ mit Hilfe des Modulationssatzes

$$\Phi_{k,k'}\left(e^{j\theta}\right) = W_{\mathrm{A}}^*\left(e^{j\left(\theta-2\pi\frac{k}{K}\right)}\right)W_{\mathrm{S}}\left(e^{j\left(\theta-2\pi\frac{k'}{K}\right)}\right). \tag{5.105}$$

Setzt man dieses Resultat in (5.101) ein, so erhält man

$$\tilde{H}_{k,k'}\left(e^{j\theta}\right) = H\left(e^{j\theta}\right)W_{\mathrm{A}}^*\left(e^{j\left(\theta-2\pi\frac{k}{K}\right)}\right)W_{\mathrm{S}}\left(e^{j\left(\theta-2\pi\frac{k'}{K}\right)}\right). \tag{5.106}$$

Da geeignete Analyse- und Synthesefenster $w_{\mathrm{A}}(l')$ und $w_{\mathrm{S}}(l')$ gewöhnlich ein sehr schmalbandiges Spektrum besitzen, wird der "Überlapp" zwischen den Funktionen $W_{\mathrm{A}}^*\left(e^{j\left(\theta-2\pi\frac{k}{K}\right)}\right)$ und $W_{\mathrm{S}}\left(e^{j\left(\theta-2\pi\frac{k'}{K}\right)}\right)$ mit wachsender Differenz $|k-k'|$ mod K geringer und die Leistung von $\tilde{H}_{k,k'}\left(e^{j\theta}\right)$ nimmt ab. Da die *DTFT* von $h_{k,k'}(m)$ wegen (5.98) durch

$$H_{k,k'}\left(e^{j\theta}\right) = \frac{1}{B}\sum_{m=0}^{B-1}\tilde{H}_{k,k'}\left(e^{j\frac{1}{B}(\theta-2\pi m)}\right). \tag{5.107}$$

ausgedrückt werden kann (siehe Kap. A.2.2), ist der Einfluss des Segmentvorschubs B auf die Kreuzbandfilter in der Regel kompliziert.

Aufgrund der aus (5.96) resultierenden Ungleichung

$$\left|\phi_{k,k'}(l)\right| \leq \sum_{l'=0}^{L_w-1}\left|w_{\mathrm{A}}(l')w_{\mathrm{S}}(l'+l)\right| \tag{5.108}$$

und der beiden Bedingungen (2.1) und (5.86) ist der Träger von $\phi_{k,k'}(l)$ durch $[-L_w+1,\ L_w-1\ ]$ gegeben. Daher sind die Kreuzbandfilter $h_{k,k'}(m'')$, welche mit Hilfe von (5.98), (5.97) und (5.78) durch

$$h_{k,k'}(m'') = \sum_{l=-L_w+1}^{L_w-1} \phi_{k,k'}(l)\, h(m''B-l) \tag{5.109}$$

$$= \sum_{l=\max(-L_w+1,\,m''B-L_h+1)}^{\min(L_w-1,\,m''B)} \phi_{k,k'}(l)\, h(m''B-l) \tag{5.110}$$

ausgedrückt werden können, im Allgemeinen bezüglich m'' nicht kausal. Der Träger ergibt sich zu $[-L_{H,\mathrm{u}}, L_H]$, wobei die Grenzen wie folgt definiert sind

$$L_{H,\mathrm{u}} := \left\lfloor \frac{L_w-1}{B} \right\rfloor \tag{5.111}$$

$$L_H := \left\lfloor \frac{L_h+L_w-2}{B} \right\rfloor. \tag{5.112}$$

Als Folge dessen treten in (5.95) bei der Summation bezüglich m'' nur endlich viele Summanden auf, d. h.

$$S(m,k) = \sum_{k'=0}^{K-1} \sum_{m''=-L_{H,\mathrm{u}}}^{L_H} X(m-m'',k')\, h_{k,k'}(m''). \tag{5.113}$$

Wird in einem letzten Schritt (5.113) in (5.79) eingesetzt, erhält man den gesuchten Zusammenhang für das Kurzzeit-Spektrum

$$Y(m,k) = \sum_{k'=0}^{K-1} \sum_{m''=-L_{H,\mathrm{u}}}^{L_H} X(m-m'',k')\, h_{k,k'}(m'') + N(m,k). \tag{5.114}$$

5.2.2. Zusammenhang im log-MEL-spektralen Bereich

Die LMSKs $y_{m,q}^{(\mathrm{s})}$ werden aus dem Kurzzeit-Leistungsspektrum des verhallten und gestörten Signals $y(l)$ gemäß

$$y_{m,q}^{(\mathrm{s})} = \ln\left\{\mathscr{Y}_{m,q}\right\} = \ln\left(\sum_{k=K_q^{(\mathrm{u})}}^{K_q^{(\mathrm{o})}} |Y(m,k)|^2\, \Lambda_q(k)\right) \tag{5.115}$$

berechnet, was durch Einsetzen von (2.4) in (2.5) ersichtlich wird. Stellt man das Kurzzeit-Leistungsspektrum von $y(l)$ mit Hilfe von (5.114) gemäß

$$|Y(m,k)|^2 = \sum_{k',k''=0}^{K-1} \sum_{m',m''=-L_{H,\mathrm{u}}}^{L_H} X(m-m',k')X^*(m-m'',k'')\, h_{k,k'}(m')h^*_{k,k''}(m'')$$

$$+ \sum_{k'=0}^{K-1} \sum_{m'=-L_{H,\mathrm{u}}}^{L_H} 2\Re\left[X(m-m',k')\, h_{k,k'}(m')N^*(m,k)\right] + |N(m,k)|^2 \tag{5.116}$$

dar, wobei $\Re[\cdot]$ den Realteil bezeichnet, so wird erkennbar, dass eine perfekte Darstellung von $y_{m,q}^{(\mathrm{s})}$ nur mit Hilfe der Kenntnis von $\left\{x_{m,q}^{(\mathrm{s})}, n_{m,q}^{(\mathrm{s})} | m \in \mathbb{Z}, q \in \{0,\ldots,Q-1\}\right\}$ sowie der Impulsantwort $h(l)$ schon deshalb nicht möglich sein kann, weil bei der Berechnung von $x_{m,q}^{(\mathrm{s})}$ und $n_{m,q}^{(\mathrm{s})}$ analog zu (5.115) jegliche Phaseninformation über die Kurzzeit-Spektren $X(m,k)$ und $N(m,k)$ verloren geht, welche zur Berechnung von (5.115) notwendig ist.

Eine mögliche Approximation von (5.116), welche nur die Kurzzeit-Leistungsspektren von $x(l)$ und $n(l)$ verwendet, ist durch

$$|Y(m,k)|^2 \approx C_{\mathrm{E}} \cdot \sum_{m'=0}^{L_H} \left|X(m-m',k)\right|^2 \left|h_{k,k}(m')\right|^2 + |N(m,k)|^2 \tag{5.117}$$

mit $C_{\mathrm{E}} \in \mathbb{R}$ gegeben, welche zudem durch die folgenden Überlegungen motiviert ist. Erstens wird der zweite Summand in (5.116) mit dem Hintergrund vernachlässigt, dass dieser unter der Annahme, dass das Störsignal $\breve{n}(l)$ mittelwertfrei und unkorreliert mit dem Sprachsignal $\breve{x}(l)$ ist, im Mittel verschwindet. Zweitens wird zur Berechnung des ersten Summanden in (5.116) der Einfluss aller Kreuzbandfilter $h_{k,k'}(m')$ bzw. $h_{k,k''}(m'')$ mit $k' \neq k$ bzw. $k'' \neq k$ ignoriert, was dadurch gerechtfertigt werden kann, dass sich deren Einfluss gemäß der Diskussion in Kap. 5.2.1 für wachsende Werte von $|k'-k|$ mod K bzw. $|k''-k|$ mod K verringert. Drittens werden im ersten Summanden von (5.116) alle Terme mit $m'' \neq m'$ fortgelassen. Diese Operation kann dadurch motiviert werden, dass die Korrelation zwischen $\breve{X}(m-m',k')$ und $\breve{X}^*(m-m'',k')$ in der Regel für wachsende Werte von $|m'-m''|$ geringer wird.

Viertens werden in (5.117) im Gegensatz zu (5.116) nur Summanden für nichtnegative Segmentindizes m' und m'' betrachtet, um einen kausalen Zusammenhang zu erhalten. Dazu ist bemerken, dass für die Merkmalsextraktion gemäß dem *ETSI-SFE*, die ja hier vordergründig betrachtet wird, die Vernachlässigung der negativen Segmentindizes nur einen sehr geringen Fehler für vernünftige Nachhallzeiten T_{60} liefert. Das hängt damit zusammen, dass in diesem Fall die Kreuzbandfilter basierend auf (5.110) durch

$$h_{k,k'}(m'') = \sum_{l=-L_w+1}^{m''B} \phi_{k,k'}(l) h(m''B-l) \quad \text{für} \quad m'' < 0 \tag{5.118}$$

berechnet werden können und der Betrag von $\phi_{k,k'}(l)$ für wachsende $|l|$ abnimmt.

Die Konstante C_{E} soll sicherstellen, dass die Approximation (5.117) erwartungstreu ist und muss dazu folgende Bedingung erfüllen

$$\mathrm{E}\left[\left|\sum_{k'=0}^{K-1}\sum_{m'=-L_{H,\mathrm{u}}}^{L_H} \breve{X}(m-m',k') h_{k,k'}(m')\right|^2\right] \stackrel{!}{=} \mathrm{E}\left[C_{\mathrm{E}} \cdot \sum_{m'=0}^{L_H} \left|\breve{X}(m-m',k)\right|^2 \left|h_{k,k}(m')\right|^2\right], \tag{5.119}$$

wobei bei gegebener Impulsantwort $h(l)$ der Erwartungswert über alle möglichen Realisierungen von $\breve{x}(l)$ zu bilden ist.

Setzt man die Approximation (5.117) des Kurzzeit-Leistungsspektrums des verhallten und

gestörten Signals $y(l)$ in (5.115) ein, so erhält man

$$y_{m,q}^{(\mathrm{s})} = \ln\left\{\sum_{k=K_q^{(\mathrm{u})}}^{K_q^{(\mathrm{o})}}\left[C_{\mathrm{E}}\cdot\sum_{m'=0}^{L_H}\left|X(m-m',k)\right|^2\left|h_{k,k}(m')\right|^2+\left|N(m,k)\right|^2\right]\Lambda_q(k)\right\} \tag{5.120}$$

$$= \ln\left\{C_{\mathrm{E}}\cdot\sum_{m'=0}^{L_H}\sum_{k=K_q^{(\mathrm{u})}}^{K_q^{(\mathrm{o})}}\left|X(m-m',k)\right|^2\left|h_{k,k}(m')\right|^2\Lambda_q(k)+\sum_{k=K_q^{(\mathrm{u})}}^{K_q^{(\mathrm{o})}}\left|N(m,k)\right|^2\Lambda_q(k)\right\}. \tag{5.121}$$

Man erkennt an dieser Stelle, dass eine weitere endgültige Approximation notwendig ist, um die log-MEL-spektralen Merkmale $y_{m,q}^{(\mathrm{s})}$ des verhallten und gestörten Sprachsignals $y(l)$ durch die log-MEL-spektralen Merkmale $x_{m,q}^{(\mathrm{s})}$ und $n_{m,q}^{(\mathrm{s})}$ des sauberen Sprachsignals $x(l)$ und des Störsignals $n(l)$ beschreiben zu können. Dazu werden die MEL-spektralen Koeffizienten $\mathscr{Y}_{m,q}$ zusätzlich dadurch angenähert, dass der in (5.121) auftretende Term $\left|h_{k,k}(m')\right|^2$ durch seinen Mittelwert über das q-te MEL-Band

$$\bar{\mathscr{H}}_{m',q} := \frac{1}{K_q^{(\mathrm{o})}-K_q^{(\mathrm{u})}+1}\sum_{k=K_q^{(\mathrm{u})}}^{K_q^{(\mathrm{o})}}\left|h_{k,k}(m')\right|^2 \tag{5.122}$$

ersetzt wird:

$$\mathscr{Y}_{m,q} \approx C_{\mathrm{E}}\cdot\sum_{m'=0}^{L_H}\bar{\mathscr{H}}_{m',q}\sum_{k=K_q^{(\mathrm{u})}}^{K_q^{(\mathrm{o})}}\left|X(m-m',k)\right|^2\Lambda_q(k)+\sum_{k=K_q^{(\mathrm{u})}}^{K_q^{(\mathrm{o})}}\left|N(m,k)\right|^2\Lambda_q(k) \tag{5.123}$$

$$= C_{\mathrm{E}}\cdot\sum_{m'=0}^{L_H}\bar{\mathscr{H}}_{m',q}\mathscr{X}_{m-m',q}+\mathscr{N}_{m,q}. \tag{5.124}$$

Wird schließlich der Fehler, der bei dieser Approximation entsteht, durch

$$\mathscr{E}_{m,q} := \mathscr{Y}_{m,q}-C_{\mathrm{E}}\cdot\sum_{m'=0}^{L_H}\bar{\mathscr{H}}_{m',q}\mathscr{X}_{m-m',q}+\mathscr{N}_{m,q} \tag{5.125}$$

definiert, ergibt sich das endgültige Beobachtungsmodell durch Einsetzen von (5.124) und (5.125) in (5.115) zu

$$y_{m,q}^{(\mathrm{s})} = \ln\left\{C_{\mathrm{E}}\cdot\sum_{m'=0}^{L_H}\bar{\mathscr{H}}_{m',q}\mathscr{X}_{m-m',q}+\mathscr{N}_{m,q}+\mathscr{E}_{m,q}\right\} \tag{5.126}$$

$$= \ln\left\{\sum_{m'=0}^{L_H}e^{x_{m-m',q}^{(\mathrm{s})}+\bar{h}_{m',q}}+e^{n_{m,q}^{(\mathrm{s})}}\right\}+v_{m,q}^{(\mathrm{s})}, \tag{5.127}$$

wobei

$$\bar{h}_{m',q} := \ln\left(C_{\mathrm{E}}\cdot\bar{\mathscr{H}}_{m',q}\right) \tag{5.128}$$

als Koeffizienten der RIA im log-MEL-spektralen Bereich angesehen werden können und

$$v_{m,q}^{(\mathrm{s})} := \ln\left\{1 + \frac{\mathscr{E}_{m,q}}{C_{\mathrm{E}} \cdot \sum\limits_{m'=0}^{L_H} \bar{\mathscr{H}}_{m',q}\,\mathscr{X}_{m-m',q} + \mathscr{N}_{m,q}}\right\}. \tag{5.129}$$

als Beobachtungsfehler interpretiert werden kann, der durch eine Prädiktion von $y_{m,q}^{(\mathrm{s})}$ basierend auf der alleinigen Kenntnis von $\left\{x_{m,q}^{(\mathrm{s})},\ldots,x_{m-L_H,q}^{(\mathrm{s})},\bar{h}_{0,q},\ldots,\bar{h}_{L_H,q},n_{m,q}^{(\mathrm{s})}\right\}$ entsteht.

Mit der Verwendung der Vektornotation und der Einführung der Beobachtungsfunktion

$$f_{\mathrm{O}} : \mathbb{R}^{[2(L_H+1)+1]Q} \to \mathbb{R}^{Q}, \quad f_{\mathrm{O}}\left(\mathbf{x}_{m:m-L_H}^{(\mathrm{s})},\bar{\mathbf{h}}_{0:L_H},\mathbf{n}_{m}^{(\mathrm{s})}\right) := \ln\left\{\sum_{m'=0}^{L_H} e^{\mathbf{x}_{m-m'}^{(\mathrm{s})}+\bar{\mathbf{h}}_{m'}} + e^{\mathbf{n}_{m}^{(\mathrm{s})}}\right\}, \tag{5.130}$$

wobei die Anwendung der mathematischen Operationen komponentenweise zu verstehen ist, lässt sich der gefundene Zusammenhang (5.127) zwischen den LMSKs kompakt gemäß

$$\mathbf{y}_{m}^{(\mathrm{s})} = f_{\mathrm{O}}\left(\mathbf{x}_{m:m-L_H}^{(\mathrm{s})},\bar{\mathbf{h}}_{0:L_H},\mathbf{n}_{m}^{(\mathrm{s})}\right) + \mathbf{v}_{m}^{(\mathrm{s})} \tag{5.131}$$

formulieren.

In Abwesenheit von Hintergrundstörungen vereinfacht sich die Beobachtungsfunktion (5.130) zu

$$\tilde{f}_{\mathrm{O}} : \mathbb{R}^{[2(L_H+1)]Q} \to \mathbb{R}^{Q}, \quad \tilde{f}_{\mathrm{O}}\left(\mathbf{x}_{m:m-L_H}^{(\mathrm{s})},\bar{\mathbf{h}}_{0:L_H}\right) := \ln\left\{\sum_{m'=0}^{L_H} e^{\mathbf{x}_{m-m'}^{(\mathrm{s})}+\bar{\mathbf{h}}_{m'}}\right\}, \tag{5.132}$$

was ersichtlich wird, indem der Grenzwert von (5.130) für $\mathbf{n}_{m}^{(\mathrm{s})} \to (-\infty,\ldots,-\infty)^T$ gebildet wird. In diesem Fall gilt entsprechend

$$\mathbf{y}_{m}^{(\mathrm{s})} = \mathbf{s}_{m}^{(\mathrm{s})} \approx \tilde{f}_{\mathrm{O}}\left(\mathbf{x}_{m:m-L_H}^{(\mathrm{s})},\bar{\mathbf{h}}_{0:L_H}\right). \tag{5.133}$$

Einen qualitativen Eindruck von der Güte dieser Approximation liefert der Vergleich der Trajektorie der wahren LMSKs-Vektoren eines beispielhaften verhallten Sprachsignals mit der entsprechenden Näherung gemäß (5.133), die jeweils in Abb. 5.5a und Abb. 5.5b dargestellt sind. Es lässt sich erkennen, dass in der approximativ berechneten Trajektorie zwar sehr feine Details nicht mehr aufgelöst werden, jedoch zumindest der grobe Verlauf korrekt dargestellt wird. Der glatte Verlauf resultiert dabei hauptsächlich aus den Näherungen (5.117) und (5.123).

Interpretation der Koeffizienten der RIA

Die in (5.128) definierten Koeffizienten der RIA $\bar{h}_{m,q}$ haben große Ähnlichkeit zu den tatsächlichen LMSKs $h_{m,q}^{(\mathrm{s})}$, welche sich gemäß (2.4) und (2.5) durch

$$h_{m,q}^{(\mathrm{s})} = \ln\left(\sum_{k=K_q^{(\mathrm{u})}}^{K_q^{(\mathrm{o})}} \Lambda_q(k)\,|H(m,k)|^2\right) \tag{5.134}$$

berechnen lassen. Dabei kann das Betragsquadrat von $H(m,k)$ mit Hilfe von (2.3) und (2.2) durch

$$|H(m,k)|^2 = \left| \sum_{l'=0}^{L_w-1} w_{\mathrm{A}}(l') h(mB+l') \cdot e^{-j\frac{2\pi}{K}kl'} \right|^2 \tag{5.135}$$

ausgedrückt werden. Die Ähnlichkeit von $h^{(\mathrm{s})}_{m,q}$ zu $\bar{h}_{m,q}$ wird erkennbar, wenn zunächst $\bar{h}_{m',q}$ unter Verwendung von (5.128) und (5.122) gemäß

$$\bar{h}_{m,q} = \ln\left(C_{\mathrm{E}} \sum_{k=K_q^{(\mathrm{u})}}^{K_q^{(\mathrm{o})}} \frac{1}{K_q^{(\mathrm{o})} - K_q^{(\mathrm{u})} + 1} \left| h_{k,k}(m) \right|^2 \right) \tag{5.136}$$

dargestellt und anschließend das Betragsquadrat von $h_{k,k}(m')$ mit Hilfe von (5.109) und (5.138) gemäß

$$\left| h_{k,k}(m') \right|^2 = \left| \sum_{p'=-L_w+1}^{L_w-1} \phi_{k,k}(-p') h(m'B+p') \right|^2 \tag{5.137}$$

geschrieben wird. Zur weiteren Umformung von (5.137) lässt sich die aus der Definition (5.96) resultierende Gleichheit

$$\phi_{k,k}(-l) = w(l) e^{-j\frac{2\pi}{K}kl} \tag{5.138}$$

ausnutzen, wobei

$$w(l) := \sum_{l'=0}^{L_w-1} w_{\mathrm{A}}(l') w_{\mathrm{S}}(l'-l) \tag{5.139}$$

eine Fensterfunktion ist, die aus der Faltung des Analysefensters $w_{\mathrm{A}}(l)$ mit dem zeitumgekehrten Synthesefenster $w_{\mathrm{S}}(-l)$ entsteht. Mit Hilfe von (5.138) erhält man schließlich

$$\left| h_{k,k}(m') \right|^2 = \left| \sum_{p'=-L_w+1}^{L_w-1} w(p') h(m'B+p') e^{-j\frac{2\pi}{K}kp'} \right|^2 . \tag{5.140}$$

Durch den Vergleich von (5.134) mit (5.136) wird ersichtlich, dass sich die Berechnung von $\bar{h}_{m,q}$ von der Berechnung von $h^{(\mathrm{s})}_{m,q}$ einerseits durch die Verwendung eines Rechteck- statt Dreickfensters zur Berechnung der Leistung für einzelne MEL-Bänder sowie der zusätzlichen Verwendung der Konstanten C_{E} in (5.136) unterscheidet. Andererseits offenbart der Vergleich von (5.135) mit (5.140) eine unterschiedliche Wahl des Analysefensters.

5.2.3. Approximation durch vereinfachtes Modell der RIA

Im Hinblick auf eine Verwendung des hergeleiteten Zusammenhanges (5.131) als Beobachtungsmodell zur BAYES'schen Merkmalsverbesserung ergeben sich in der Praxis mehrere Schwierigkeiten. Für ein Szenario, in dem die RIA vom Sprecher zum Mikrophon unbekannt ist, besteht das grundsätzliche Problem der Berechnung der Koeffizienten (5.128).

Zwar ist es möglich, die RIA aus dem aufgenommenen Mikrophonsignal zu schätzen und anschließend für die Berechnung der Koeffizienten (5.128) zu verwenden. Dabei wird jedoch die Schätzung durch die Tatsache erschwert, dass die RIA in der Regel, bedingt u.a. durch Bewegungen des Sprechers oder Änderungen der Temperatur und Feuchtigkeit innerhalb des Raumes, zeitvariant ist. Die zeitlichen Änderungen betreffen dabei häufig nur die feine Struktur, wobei die Einhüllende ihre Form beibehält (siehe Kap. 2.3). Außerdem besitzt die RIA sehr viele Koeffizienten, so dass eine zuverlässige Schätzung insbesondere bei Räumen mit größeren Nachhallzeiten im Allgemeinen nicht trivial ist.

Motiviert durch diese Überlegungen wird das stark vereinfachte Modell der RIA

$$\breve{h}(l) = \sigma_h \cdot \breve{v}_h(l) \cdot \chi_h(l) \cdot e^{-\frac{l}{\tau_h}}, \tag{5.141}$$

verwendet, welches bereits in [Pol88] eingeführt wurde. Dabei bezeichnet $\breve{v}_h(l)$ einen mittelwertfreien weißen GAUSS'schen Zufallsprozess, dessen Autokorrelationsfunktion durch

$$\mathrm{E}\left[\breve{v}_h(l)\breve{v}_h(l')\right] = \delta\left(l - l'\right) \quad \text{für} \quad l, l' \in \mathbb{Z} \tag{5.142}$$

gegeben ist und der durch die Zufälligkeit der Reflexionen der akustischen Wellen an Oberflächen motiviert ist. Der Faktor $e^{-\frac{l}{\tau_h}}$ erzeugt eine exponentiell abklingende Einhüllende, wobei die Abklingkonstante τ_h wie folgt mit der mittleren Nachhallzeit T_{60} und der Abtastdauer T_A zusammenhängt (siehe Kap. A.2.3 im Anhang):

$$\tau_h = \frac{T_{60}}{3\ln(10) \cdot T_\mathrm{A}}. \tag{5.143}$$

Die Funktion

$$\chi_h(l) := \begin{cases} 1 & \text{für } 0 \leq l \leq L_h - 1 \\ 0 & \text{sonst} \end{cases} \tag{5.144}$$

kann als Indikatorfunktion von $h(l)$ angesehen werden und sorgt dafür, dass die RIA kausal wird und eine endliche Länge L_h aufweist. Der Skalierungsfaktor σ_h bestimmt die mittlere Leistung der RIA, welche sich durch Anwendung der geometrischen Summe

$$\sum_{l=0}^{L-1} x^l = \frac{x^L - 1}{x - 1} \quad \text{für } x \in \mathbb{C} \setminus \{1\},\ L \in \mathbb{N} \tag{5.145}$$

unter Berücksichtigung von (5.142) durch

$$\mathrm{E}\left[\sum_{l=0}^{L_h-1} \breve{h}^2(l)\right] = \sigma_h^2 \sum_{l=0}^{L_h-1} e^{-\frac{2l}{\tau_h}} = \sigma_h^2 \cdot \frac{e^{-\frac{2L_h}{\tau_h}} - 1}{e^{-\frac{2}{\tau_h}} - 1} \tag{5.146}$$

berechnen lässt.

Im Folgenden soll angenommen werden, dass keine detaillierte Kenntnis der RIA vorliegt, jedoch lediglich bekannt ist, dass diese eine Realisierung des in (5.141) definierten Zufallsprozesses darstellt, wobei die beiden Parameter τ_h und σ_h gegeben sind. Bedingt durch diese Annahme stellt nun jeder Koeffizient der RIA im log-MEL-spektralen Bereich $\bar{h}_{m',q}$,

welcher sich gemäß (5.128) aus der der RIA berechnen lässt, ebenfalls eine Realisierung einer Zufallsvariable $\breve{\breve{h}}_{m',q}$ dar. Um in dieser Situation zu einer sinnvollen Wahl für die Koeffizienten $\bar{h}_{m',q}$ zur Verwendung im Beobachtungsmodell (5.131) zu gelangen, erscheint es sinnvoll, diese Koeffizienten durch ihren Erwartungswert

$$\mu_{\breve{\bar{h}}_{m',q}} := \mathrm{E}\left[\breve{\breve{h}}_{m',q}\right]. \tag{5.147}$$

basierend auf dem Modell der RIA zu ersetzen. Eine analytische Berechnung dieses Erwartungswertes ist wegen der auftretenden Logarithmusoperation sehr aufwendig. Sie wird jedoch stark durch die approximative Annahme vereinfacht, dass die Verteilungsdichtefunktion von $\breve{h}^{(\mathrm{s})}_{m',q}$ durch eine GAUSS-Verteilung mit dem Mittelwert $\mu_{\breve{\bar{h}}_{m',q}}$ und der Varianz $\sigma^2_{\breve{\bar{h}}_{m',q}}$ beschrieben werden kann, d.h.

$$p_{\breve{\bar{h}}_{m',q}}\left(\bar{h}_{m',q}\right) = \mathcal{N}\left(\bar{h}_{m',q}; \mu_{\breve{\bar{h}}_{m',q}}, \sigma^2_{\breve{\bar{h}}_{m',q}}\right). \tag{5.148}$$

Basierend auf dieser Annahme und (5.128) sind die MEL-spektalen Koeffizienten der RIA $\breve{\breve{\mathscr{H}}}_{m',q}$ log-normalverteilt, wobei sich insbesondere der Erwartungswert (5.147) aus dem Erwartungswert und der Varianz

$$\mu_{\breve{\mathscr{H}}_{m',q}} := \mathrm{E}\left[\breve{\breve{\mathscr{H}}}_{m',q}\right] \tag{5.149}$$

$$\sigma^2_{\breve{\mathscr{H}}_{m',q}} := \mathrm{E}\left[\left(\breve{\breve{\mathscr{H}}}_{m',q} - \mu_{\breve{\mathscr{H}}_{m',q}}\right)^2\right] \tag{5.150}$$

gemäß

$$\mu_{\breve{\bar{h}}_{m',q}} = \mu_{\breve{\bar{h}}_{m',q}}(\tau_h, \sigma_h) = \frac{1}{2}\ln\left\{\frac{\left[\mu_{\breve{\mathscr{H}}_{m',q}}\right]^4}{\sigma^2_{\breve{\mathscr{H}}_{m',q}} + \left[\mu_{\breve{\mathscr{H}}_{m',q}}\right]^2}\right\}. \tag{5.151}$$

darstellen lässt [AB57]. Obwohl die Berechnung der Varianz im Sinne des Beobachtungsmodells nicht notwendig ist, sei zur Vollständigkeit bemerkt, dass diese durch

$$\sigma^2_{\breve{\bar{h}}_{m',q}} := \mathrm{E}\left[\left(\breve{\breve{h}}_{m',q} - \mu_{\breve{\bar{h}}_{m',q}}\right)^2\right] = \ln\left\{\frac{\sigma^2_{\breve{\mathscr{H}}_{m',q}}}{\left[\mu_{\breve{\mathscr{H}}_{m',q}}\right]^2} + 1\right\} \tag{5.152}$$

gegeben ist. In Kap. A.2.4 im Anhang wird gezeigt, dass der Mittelwert (5.149) und die Varianz (5.150) durch

$$\mu_{\breve{\mathscr{H}}_{m',q}} = \sum_{p'=-L_w+1}^{L_w-1} \delta^2_{m',p',0} \tag{5.153}$$

$$\sigma^2_{\breve{\mathscr{H}}_{m',q}} = \frac{1}{\left(K_q^{(\mathrm{o})} - K_q^{(\mathrm{u})} + 1\right)^2} \sum_{k,k'=K_q^{(\mathrm{u})}}^{K_q^{(\mathrm{o})}} \left(\left| \sum_{p'=-L_w+1}^{L_w-1} \delta^2_{m',p',\frac{k+k'}{2}} \right|^2 + \left| \sum_{p'=-L_w+1}^{L_w-1} \delta^2_{m',p',\frac{k-k'}{2}} \right|^2 \right) \tag{5.154}$$

berechnet werden können, wobei

$$\delta_{m',p',k} := \sigma_h \cdot \chi_h(m'B+p') \cdot e^{-\frac{m'B+p'}{\tau_h}} w(p') e^{-j\frac{2\pi}{K}kp'}. \tag{5.155}$$

Dass diese Näherung sinnvoll ist, zeigt der Vergleich der mit Hilfe von MONTE-CARLO-Simulationen erzeugten normierten Histogramme mit den entsprechenden analytischen GAUSS-förmigen Approximationen in Abb. 5.2 und Abb. 5.3 für zwei beispielhafte Nachhallzeiten von $T_{60} = 0{,}1\,\mathrm{s}$ und $T_{60} = 0{,}8\,\mathrm{s}$. Dabei wird erkennbar, dass die analytisch berechnete Approximation in der Regel umso besser ist, je größer der Index q des MEL-Bandes ist. Zudem kann beobachtet werden, dass die Varianz für wachsende Indizes q sinkt.

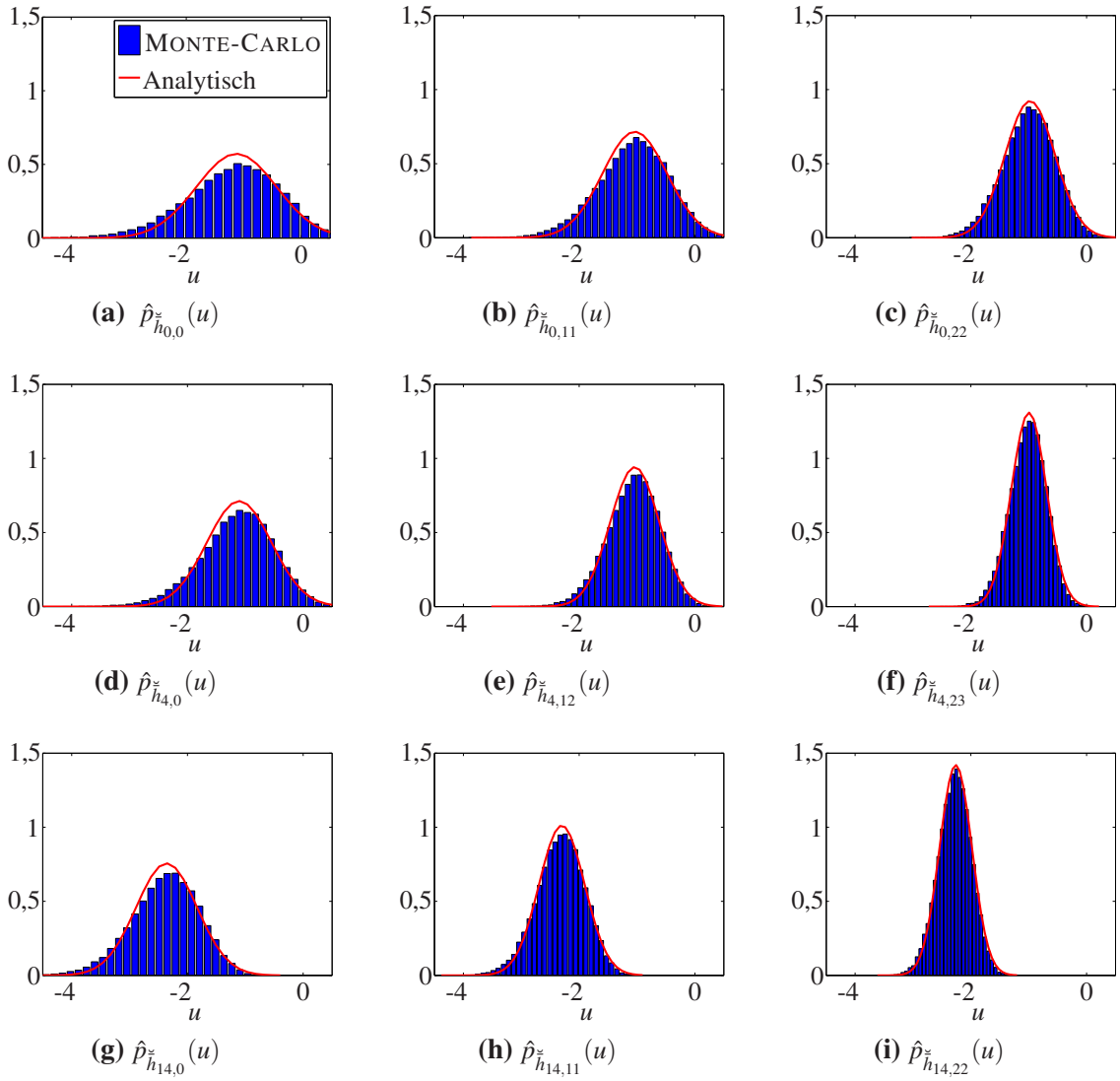
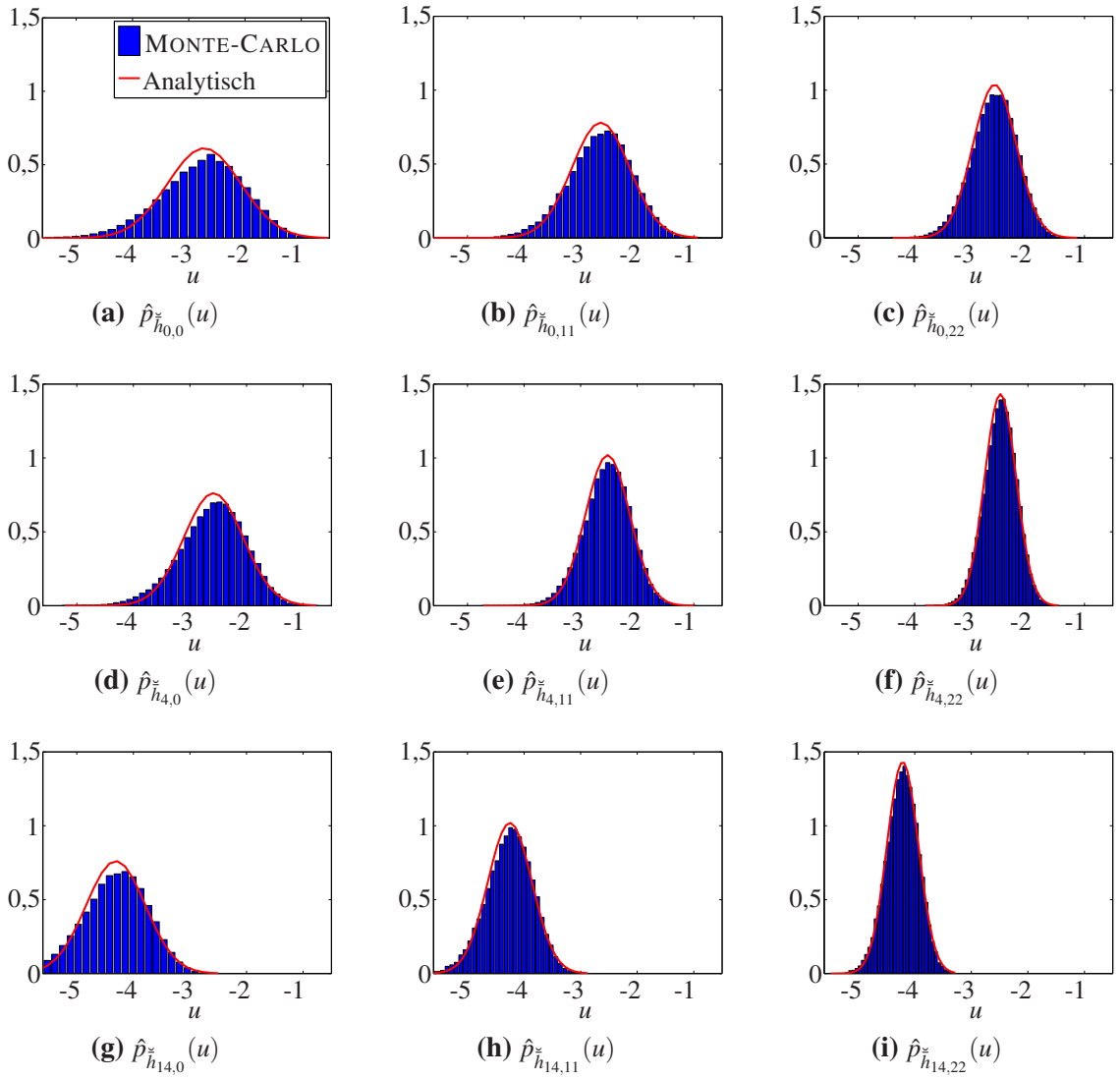


Abbildung 5.2.: *Approximationen $\hat{p}_{\breve{h}_{m,q}}(u)$ der Verteilungsdichtefunktionen der log-*MEL*-spektralen Repräsentationen der RIA $\bar{h}_{m,q}$ für $m \in \{0,4,14\}$ und $q \in \{0,11,22\}$ durch normierte Histogramme, resultierend aus* MONTE-CARLO*-Simulationen (blau) einerseits, sowie aus einer analytischen Darstellung (rot) andererseits, für eine Nachhallzeit von $T_{60} = 0{,}1\,\mathrm{s}$.*

Nachdem nun das Prinzip der Berechnung der log-MEL-spektralen Repräsentation der RIA basierend auf dem Modell (5.141) mit gegebenen Parametern τ_h sowie σ_h erläutert wurde, soll das Augenmerk auf den Aspekt gerichtet werden, wie die Länge der RIA L_h sowie die Leistungskompensationskonstante C_{E} berechnet werden können.



(a) $\hat{p}_{\breve{h}_{0,0}}(u)$ **(b)** $\hat{p}_{\breve{h}_{0,11}}(u)$ **(c)** $\hat{p}_{\breve{h}_{0,22}}(u)$

(d) $\hat{p}_{\breve{h}_{4,0}}(u)$ **(e)** $\hat{p}_{\breve{h}_{4,11}}(u)$ **(f)** $\hat{p}_{\breve{h}_{4,22}}(u)$

(g) $\hat{p}_{\breve{h}_{14,0}}(u)$ **(h)** $\hat{p}_{\breve{h}_{14,11}}(u)$ **(i)** $\hat{p}_{\breve{h}_{14,22}}(u)$

Abbildung 5.3.: *Approximationen $\hat{p}_{\breve{h}_{m,q}}(u)$ der Verteilungsdichtefunktionen der log-MEL-spektralen Repräsentationen der RIA $\bar{h}_{m,q}$ für $m \in \{0,4,14\}$ und $q \in \{0,11,22\}$ durch normierte Histogramme, resultierend aus MONTE-CARLO-Simulationen (blau) einerseits, sowie aus einer analytischen Darstellung (rot) andererseits, für eine Nachhallzeit von $T_{60} = 0{,}8\,\mathrm{s}$.*

Wahl der Länge L_h der RIA

Zunächst ist es wichtig festzustellen, dass eine sinnvolle Wahl der Länge L_h in irgendeiner Art und Weise von der Nachhallzeit T_{60} bzw. von der Abklingkonstanten τ_h abhängen sollte.

Ein mögliches Kriterium für eine solche Wahl besteht darin, eine von τ_h unabhängige Konstante $\varepsilon_h < 1$ vorzugeben, welche die Güte der Modellierung basierend auf dem relativen Leistungsverhältnis

$$r(L_h) := \frac{\mathrm{E}\left[\sum_{l'=0}^{L_h-1} \breve{h}^2(l')\right]}{\mathrm{E}\left[\sum_{l'=0}^{\infty} \breve{h}^2(l')\right]} = 1 - \frac{\mathrm{E}\left[\sum_{l'=L_h}^{\infty} \breve{h}^2(l')\right]}{\mathrm{E}\left[\sum_{l'=0}^{\infty} \breve{h}^2(l')\right]} = 1 - e^{-\frac{2L_h}{\tau_h}} \tag{5.156}$$

zwischen der abgeschnittenen und der nicht abgeschnittenen RIA durch

$$r(L_h) > 1 - \varepsilon_h \tag{5.157}$$

beschreibt. Für die Umformungen in (5.156) wurden (2.21) sowie (A.104) verwendet. In anderen Worten ausgedrückt bedeutet (5.157), dass die relative Leistung des abgeschnittenen Anteils der RIA, welche gleich $1 - r(L_h)$ ist, geringer als ε_h sein muss. Die Länge der RIA kann dann unter Einhaltung der Nebenbedingung (5.157) minimiert werden, was schließlich in

$$\hat{L}_h(\tau_h) := \underset{L_h}{\operatorname{argmin}}\, r(L_h) \quad \text{unter der Nebenbedingung (5.157)} \tag{5.158}$$

$$= \left\lceil -\frac{\tau_h}{2} \ln(\varepsilon_h) \right\rceil \tag{5.159}$$

resultiert.

Wahl der Leistungskompensationskonstante C_E

Zur Erinnerung sei noch einmal erwähnt, dass die Leistungskompensationskonstante C_E dazu dient, die Vernachlässigung von Termen zur vereinfachten Berechnung des Kurzzeit-Leistungsspektrums des gestörten und verhallten Signals $y(l)$ gemäß (5.117) zu kompensieren. Sie kann theoretisch mit Hilfe der Bedingung (5.119) bestimmt werden, wobei der Erwartungswert bei gegebener RIA über alle möglichen Realisierungen von $x(l)$ zu bilden ist. Für den Fall, dass die RIA sich gemäß dem stochastischen Modell (5.141) verhält, ist es möglich, den Erwartungswert zusätzlich über alle möglichen Realisierungen von $h(l)$ zu betrachten, wobei die Parameter τ_h und σ_h deterministische Größen sind.

Um die Erwartungswertbildung bezüglich $x(l)$ überhaupt handhabbar zu gestalten, soll weiterhin angenommen werden, dass es sich dabei um einen weißen GAUSS'schen Zufallsprozess handelt. Unter dieser Annahme kann gezeigt werden (siehe Kap. A.2.5 im Anhang), dass sich die Konstante C_E aus dem Quotienten

$$C_\mathrm{E} = \frac{C_\mathrm{Z}}{C_\mathrm{N}} \tag{5.160}$$

ergibt, wobei der Zähler und Nenner durch

$$C_\mathrm{Z} := K^2 \sum_{m',m''=-L_{H,\mathrm{u}}}^{L_H} \sum_{l=0}^{L_w-1} w_\mathrm{A}(l) w_\mathrm{S}(l) w_\mathrm{A}(l + (m'' - m')B) w_\mathrm{S}(l + (m'' - m')B)$$
$$\cdot \sum_{l'=-L_w+1}^{L_w-1} \chi_h(m'B - l') e^{-\frac{2(m'B-l')}{\tau_h}} w_\mathrm{A}^2(-l' + l) \tag{5.161}$$

und

$$C_{\mathrm{N}} := \left(\sum_{l=0}^{L_w-1} w_{\mathrm{A}}^2(l) \right) \left[\sum_{m'=0}^{L_H} \sum_{l'=-L_w+1}^{L_w-1} \left(\sum_{p''=0}^{L_w-1} w_{\mathrm{A}}(p'') w_{\mathrm{S}}(p''+l') \right)^2 \chi_h(m'B-l') e^{-\frac{2(m'B-l')}{\tau_h}} \right] \tag{5.162}$$

definiert sind. Sie hängt also nur von den bei der Merkmalsextraktion verwendeten Parametern sowie der Abklingkonstante τ_h ab.

Schätzung der RIA-Parameter

Das Modell der RIA (5.141) wurde mit der Motivation eingeführt, dass es nur durch zwei Parameter vollständig beschrieben ist, welche in einer dem automatischen Spracherkenner unbekannten Umgebung in der Regel deutlich einfacher aus dem eingehenden Mikrophonsignal zu schätzen sind als die gesamter RIA selbst.

Für den Spezialfall, dass keine Störung während der Spracherkennung vorhanden ist, existieren in der Literatur hauptsächlich zwei Ansätze zur blinden Schätzung der Nachhallzeit T_{60}. Bei den Verfahren basierend auf dem ***M**aximum **L**ikelihood* (*ML*)-Prinzip [RJW+03, RJO04] wird versucht, die Abklingkonstante τ_h, welche über (5.143) mit der Nachhallzeit verknüpft ist, derart zu bestimmen, dass kurze Signalausschnitte, welche vorwiegend den Übergang zwischen der Sprache und den Sprachpausen darstellen, durch das Modell der RIA bestmöglich beschrieben werden. Aus dem Histogramm der aus vielen Signalausschnitten resultierenden Schätzungen wird anschließend die gesuchte Abklingkonstante beispielsweise durch die Suche des ersten lokalen Maximums oder des 10 %-Quantils bestimmt. Hingegen wird in [WHN08] ein etwas anderer Ansatz verfolgt. Zunächst wird auch hier ein Histogramm aus Schätzungen von Abklingkonstanten geschätzt, was jedoch durch lineare Regression aus dem logarithmierten Kurzzeit-Leistungsspektrum des verhallten Signal bestimmt wird. Die endgültige Schätzung der Abklingkonstante basiert auf einem beobachteten nichtlinearen Zusammenhang zwischen der Schiefe des Histogramms und der Nachhallzeit. Für den allgemeinen Fall, bei dem eine (nicht zu starke) Störung $n(l) \neq 0$ präsent ist, kann eine Schätzung im Prinzip mit den gleichen Methoden erfolgen. Jedoch muss zuvor eine Entstörung des Signals beispielsweise mit Hilfe von spektraler Subtraktion oder eines WIENER Filters [VM06] erfolgen.

Der Parameter σ_h beschreibt im Wesentlichen den relativen Einfluss der Raumimpulsantwort auf die Leistung des verhallten Signals $\breve{s}(l)$. Aufgrund des instationären Charakters des sauberen Sprachsignals $\breve{x}(l)$ (und in manchen Situationen des Störsignals $\breve{n}(l)$) ist eine blinde Schätzung nicht trivial. Anstatt an dieser Stelle eine detaillierte Methode zu seiner Schätzung zu gegeben, soll hier nur das Prinzip unter der vereinfachten Annahme beschrieben werden, dass sowohl das saubere Sprachsignal $\breve{x}(l)$ als auch das Störsignal $\breve{n}(l)$ durch stationäre Zufallsprozesse mit den Leistungen $\sigma_x^2 := \mathrm{E}\left[\breve{x}^2(l)\right]$ und $\sigma_n^2 := \mathrm{E}\left[\breve{n}^2(l)\right]$ gegeben sind. Dann lässt sich die Leistung des verhallten und gestörten Signals $\breve{y}(l)$ mit Hilfe der Annahme, dass das saubere Sprachsignal $\breve{x}(l)$ und das Störsignal $\breve{n}(l)$ miteinander unkorreliert sind, und des

Modells für die RIA (5.141) gemäß

$$\sigma_y^2 := \mathrm{E}\left[\breve{y}^2(l)\right] = \mathrm{E}\left[\breve{s}^2(l)\right] + \mathrm{E}\left[\breve{n}^2(l)\right] \tag{5.163}$$

$$= \sigma_s^2 + \sigma_n^2 \tag{5.164}$$

$$= \mathrm{E}\left[\sum_{p'=0}^{L_h-1}\sum_{p''=0}^{L_h-1} \breve{h}(p')\breve{h}(p'')\breve{x}(l-p')\breve{x}(l-p'')\right] + \sigma_n^2 \tag{5.165}$$

$$= \sum_{p'=0}^{L_h-1}\sum_{p''=0}^{L_h-1} \mathrm{E}\left[\breve{h}(p')\breve{h}(p'')\right]\mathrm{E}\left[\breve{x}(l-p')\breve{x}(l-p'')\right] + \sigma_n^2 \tag{5.166}$$

$$= \sum_{p'=0}^{L_h-1} \mathrm{E}\left[\breve{h}^2(p')\right]\mathrm{E}\left[\breve{x}^2(l-p')\right] + \sigma_n^2 \tag{5.167}$$

$$= \sigma_x^2 \cdot \mathrm{E}\left[\sum_{p'=0}^{L_h-1} \breve{h}^2(p')\right] + \sigma_n^2 \tag{5.168}$$

$$= \sigma_x^2 \left(\sigma_h^2 \cdot \frac{e^{-\frac{2L_h}{\tau_h}} - 1}{e^{-\frac{2}{\tau_h}} - 1}\right) + \sigma_n^2 \tag{5.169}$$

ausdrücken, wobei σ_s^2 in (5.164) die Leistung des verhallten Sprachsignals $\breve{s}(l)$ bezeichnet und für die Umformung von (5.168) nach (5.169) das Resultat (5.146) verwendet wurde. Daraus ergibt sich der gesuchte Parameter σ_h zu

$$\sigma_h = \sqrt{\frac{\left(\sigma_y^2 - \sigma_n^2\right)}{\sigma_x^2} \cdot \frac{\left(e^{-\frac{2}{\tau_h}} - 1\right)}{\left(e^{-\frac{2L_h}{\tau_h}} - 1\right)}}. \tag{5.170}$$

Seine Schätzung erfordert daher die Schätzung der Leistungen der Signale $\breve{y}(l)$ und $\breve{n}(l)$, wenn man annimmt, dass die Leistung des zugrunde liegenden sauberen Sprachsignals $\breve{x}(l)$ bekannt ist.

In der Praxis werden die auftretenden Zufallsprozesse in der Regel instationär sein, sodass eine Approximation der Leistungen beispielsweise durch die Berechnung von gleitenden Mittelwerten vorgenommen werden muss. Um zwischen Signalausschnitten mit und ohne Sprachaktivität unterscheiden zu können, kann eine Sprachaktivitätsdetektion eingesetzt werden.

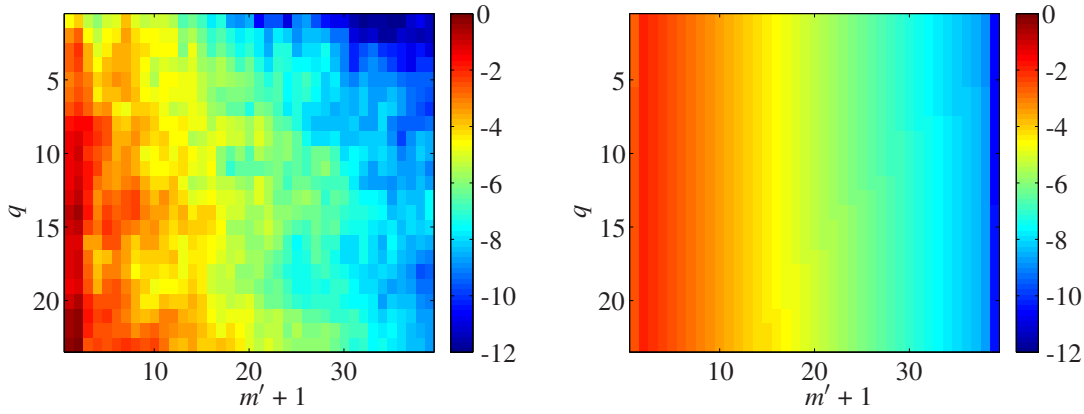
Anstelle der Schätzung des Parameters σ_h kann auch äquivalent dazu das verhallte und gestörte Sprachsignal $\breve{y}(l)$ so normiert werden, dass $\sigma_s^2 = \sigma_x^2$ gilt, und der Parameter σ_h dann einfach zu

$$\sigma_h = \sqrt{\frac{\left(e^{-\frac{2}{\tau_h}} - 1\right)}{\left(e^{-\frac{2L_h}{\tau_h}} - 1\right)}} \tag{5.171}$$

gesetzt werden.

Defizite des RIA-Modells

Die stark vereinfachte Charakterisierung der RIA durch nur zwei Parameter bringt natürlich nicht nur Vorteile mit sich. Strikt genommen würde eine derartige Beschreibung nur für den späten Nachhall der RIA zutreffen, der durch ein vollkommen diffuses Schallfeld erzeugt wird, bei dem die Reflexionen aus jeder Richtung mit derselben Wahrscheinlichkeit und Intensität auf das Mikrophon einfallen. Diese Bedingung wird in den meisten Anwendungen im Allgemeinen nicht zutreffen, sodass (besonders) die frühen Reflexionen bedingt durch die Geometrie des Raumes Korrelationen untereinander aufweisen werden. Aufgrund dessen besitzen RIAs typischerweise einen Abklang, der durch zwei unterschiedliche Abklingkonstanten gekennzeichnet ist [Sch65]. Eine weitere Tatsache, die durch das Modell (5.141) vernachlässigt wird, ist die unterschiedliche Art der Reflexion und Dämpfung von akustischen Wellen unterschiedlicher Frequenzen, welche eine Frequenzabhängigkeit des Energieabklangs der RIA mit sich bringt [Kut04]. In Abb. 5.4 werden beide Aspekte durch die visuelle Darstellung der Koeffizienten der RIA aus Abb. 2.3 veranschaulicht, wobei die wahren Koeffizienten in Abb. 5.4a ihrer Approximation in Abb. 5.4b gemäß (5.151) mit $\hat{T}_{60} = 0{,}75\,\mathrm{s}$ gegenüber gestellt sind.



(a) *Wahre Koeffizienten der RIA* $\bar{h}_{m',q}$

(b) *Approximative Koeffizienten der RIA* $\mu_{\breve{h}_{m',q}}$ *berechnet gemäß* (5.151) *mit* $\hat{T}_{60} = 0{,}75\,\mathrm{s}$

Abbildung 5.4.: *Log-*MEL*-spektrale Repräsentation der RIA aus Abb. 2.3, wobei* m' *den Segmentindex innerhalb der RIA und* q *den Index des* MEL*-Bandes bezeichnet.*

Obwohl es prinzipiell möglich wäre, ähnlich wie in [WSNK09] das Modell der RIA (5.141) derart zu verfeinern, dass die angesprochenen Eigenschaften der RIA mit erfasst werden, wird in dieser Arbeit davon abgesehen. Der Grund liegt in der mit der Verfeinerung des Modells einhergehenden steigenden Komplexität, welche sehr wahrscheinlich die Genauigkeit der blinden Schätzung der entsprechenden Modellparameter negativ beeinflussen würde.

5.2.4. Rekursives Beobachtungsmodell

Für große Nachhallzeiten T_{60} wird der Wert von L_H groß, so dass in der Beobachtungsfunktion (5.130) sehr viele Exponentialterme ausgewertet werden müssen. Außerdem sind bei der Merkmalsverbesserung die Werte der LMSKs des sauberen Sprachsignals $x_{m,q}^{(\mathrm{s})}$ natürlich unbekannt, so dass Schätzwerte eingesetzt werden müssen und der dadurch entstehende Fehler zusätzlich berücksichtigt werden muss. Mit der Motivation der Lösung beider Probleme wird im Folgenden ein rekursives Beobachtungsmodell hergeleitet, welches auf dem vereinfachten Modell der RIA (5.141) basiert.

Die Grundlage besteht in einem approximativ rekursiven Zusammenhang zwischen der Leistung von Band-zu-Band-Filtern mit verschiedenen Segmentindizes, welcher sich aus der in Kap. A.2.4 im Anhang hergeleiteten Beziehung (A.114) ergibt:

$$\mathrm{E}\left[\left|\breve{h}_{k,k}(m'+L_\mathrm{R})\right|^2\right] = \sum_{p'=-L_w+1}^{L_w-1} \delta^2_{m'+L_\mathrm{R},p',0} \tag{5.172}$$

$$= \sum_{p'=-L_w+1}^{L_w-1} \sigma_h^2 \cdot \chi_h\big(\big(m'+L_\mathrm{R}\big)B+p'\big) \cdot e^{-\frac{2\left[\left(m'+L_\mathrm{R}\right)B+p'\right]}{\tau_h}} \cdot w^2(p') \tag{5.173}$$

$$\approx e^{-\frac{2L_\mathrm{R}B}{\tau_h}} \cdot \mathrm{E}\left[\left|\breve{h}_{k,k}(m')\right|^2\right] \quad \forall m', L_\mathrm{R} \in \mathbb{N}_0. \tag{5.174}$$

Dabei ist die Approximation nur durch die zeitliche Begrenzung der RIA, welche durch ihre Indikatorfunktion $\chi_h(l)$ beschrieben wird, begründet, so dass unter der Annahme der Gültigkeit des RIA-Modells (5.141) die Rekursion (5.174) für $\frac{L_w-1}{B} \leq m'+L_\mathrm{R} \leq \frac{L_h-L_w}{B}$ sogar exakt ist.

Sie lässt sich sich verwenden, um einen approximativen rekursiven Ausdruck für den Erwartungswert des Leistungsspektrums des verhallten und gestörten Sprachsignals $y(l)$ bezüglich der RIA zu finden, welcher mit Hilfe von (5.117) und der Berücksichtigung der Tatsache, dass $h_{k,k}(m') = 0$ für $m' > L_H$ gilt, zunächst durch

$$\mathrm{E}_{\breve{h}(l)}\left[\left|\breve{Y}(m,k)\right|^2\right] \approx C_\mathrm{E} \cdot \sum_{m'=0}^{L_H} \left|X(m-m',k)\right|^2 \mathrm{E}\left[\left|\breve{h}_{k,k}(m')\right|^2\right] + |N(m,k)|^2 \tag{5.175}$$

$$= C_\mathrm{E} \cdot \left(\sum_{m'=0}^{L_\mathrm{R}-1} \left|X(m-m',k)\right|^2 \mathrm{E}\left[\left|\breve{h}_{k,k}(m')\right|^2\right] + \sum_{m'=L_\mathrm{R}}^{L_H} \left|X(m-m',k)\right|^2 \mathrm{E}\left[\left|\breve{h}_{k,k}(m')\right|^2\right] \right) + |N(m,k)|^2 \tag{5.176}$$

$$= C_\mathrm{E} \cdot \left(\sum_{m'=0}^{L_\mathrm{R}-1} \left|X(m-m',k)\right|^2 \mathrm{E}\left[\left|\breve{h}_{k,k}(m')\right|^2\right] + \sum_{m'=0}^{L_H} \left|X(m-m'-L_\mathrm{R},k)\right|^2 \mathrm{E}\left[\left|\breve{h}_{k,k}(m'+L_\mathrm{R})\right|^2\right] \right) + |N(m,k)|^2 \tag{5.177}$$

für $1 \leq L_{\mathrm{R}} \leq L_H$ schreiben lässt. Setzt man nun die Approximation (5.174) in (5.177) ein und verwendet die aus (5.175) resultierende Approximation

$$\begin{aligned} & C_{\mathrm{E}} \sum_{m'=0}^{L_H} \left|X(m-m'-L_{\mathrm{R}},k)\right|^2 \mathrm{E}\left[\left|\breve{h}_{k,k}(m')\right|^2\right] \\ \approx & \max\left\{\mathrm{E}_{\breve{h}(l)}\left[\left|\breve{Y}(m-L_{\mathrm{R}},k)\right|^2\right] - |N(m-L_{\mathrm{R}},k)|^2, 0\right\}, \end{aligned} \tag{5.178}$$

so ergibt sich

$$\begin{aligned} \mathrm{E}_{\breve{h}(l)}\left[\left|\breve{Y}(m,k)\right|^2\right] \approx & C_{\mathrm{E}} \left(\sum_{m'=0}^{L_{\mathrm{R}}-1} \left|X(m-m',k)\right|^2 \mathrm{E}\left[\left|\breve{h}_{k,k}(m')\right|^2\right] \right. \\ & \left. + e^{-\frac{2L_{\mathrm{R}}B}{\tau_h}} \sum_{m'=0}^{L_H} \left|X(m-m'-L_{\mathrm{R}},k)\right|^2 \mathrm{E}\left[\left|\breve{h}_{k,k}(m')\right|^2\right] \right) + |N(m,k)|^2 \end{aligned} \tag{5.179}$$

$$\begin{aligned} \approx & C_{\mathrm{E}} \sum_{m'=0}^{L_{\mathrm{R}}-1} \left|X(m-m',k)\right|^2 \mathrm{E}\left[\left|\breve{h}_{k,k}(m')\right|^2\right] + |N(m,k)|^2 \\ & + e^{-\frac{2L_{\mathrm{R}}B}{\tau_h}} \cdot \max\left\{\mathrm{E}_{\breve{h}(l)}\left[\left|\breve{Y}(m-L_{\mathrm{R}},k)\right|^2\right] - |N(m-L_{\mathrm{R}},k)|^2, 0\right\}. \end{aligned} \tag{5.180}$$

Die Maximumbildung in (5.178) ist dadurch bedingt, dass der zu approximierende Ausdruck stets nichtnegativ sein muss.

Motiviert durch die rekursive Approximation (5.180) lässt sich direkt eine entsprechende Beziehung zwischen den MEL-spektralen Koeffizienten finden

$$\mathscr{Y}_{m,q} \approx C_{\mathrm{E}} \cdot \sum_{m'=0}^{L_{\mathrm{R}}-1} \bar{\mathscr{H}}_{m',q} \mathscr{X}_{m-m',q} + e^{-\frac{2L_{\mathrm{R}}B}{\tau_h}} \cdot \max\left\{\mathscr{Y}_{m-L_{\mathrm{R}},q} - \mathscr{N}_{m-L_{\mathrm{R}},q}, 0\right\} + \mathscr{N}_{m,q}, \tag{5.181}$$

wobei jetzt der Erwartungswert weggelassen wurde. Definiert man den mit dieser Approximation verbundenen Fehler durch

$$\mathscr{E}^{(\mathrm{R})}_{m,L_{\mathrm{R}},q} := \mathscr{Y}_{m,q} - C_{\mathrm{E}} \cdot \sum_{m'=0}^{L_{\mathrm{R}}-1} \bar{\mathscr{H}}_{m',q} \mathscr{X}_{m-m',q} - e^{-\frac{2L_{\mathrm{R}}B}{\tau_h}} \cdot \max\left\{\mathscr{Y}_{m-L_{\mathrm{R}},q} - \mathscr{N}_{m-L_{\mathrm{R}},q}, 0\right\} - \mathscr{N}_{m,q}, \tag{5.182}$$

so gelangt man zum gewünschten Ausdruck für die LMSKs

$$y^{(\mathrm{s})}_{m,q} = \ln\left\{ \sum_{m'=0}^{L_{\mathrm{R}}-1} e^{x^{(\mathrm{s})}_{m-m',q} + \bar{h}_{m',q}} + e^{-\frac{2L_{\mathrm{R}}B}{\tau_h}} \cdot \max\left[e^{y^{(\mathrm{s})}_{m-L_{\mathrm{R}},q}} - e^{n^{(\mathrm{s})}_{m-L_{\mathrm{R}},q}}, 0\right] + e^{n^{(\mathrm{s})}_{m,q}} \right\} + v^{(\mathrm{s,R})}_{m,L_{\mathrm{R}},q} \tag{5.183}$$

mit

$$v^{(\mathrm{s,R})}_{m,L_{\mathrm{R}},q} := \ln\left\{ 1 + \frac{\mathscr{E}^{(\mathrm{R})}_{m,L_{\mathrm{R}},q}}{\sum_{m'=0}^{L_{\mathrm{R}}-1} e^{x^{(\mathrm{s})}_{m-m',q} + \bar{h}_{m',q}} + e^{-\frac{2L_{\mathrm{R}}B}{\tau_h}} \cdot \max\left[e^{y^{(\mathrm{s})}_{m-L_{\mathrm{R}},q}} - e^{n^{(\mathrm{s})}_{m-L_{\mathrm{R}},q}}, 0\right] + e^{n^{(\mathrm{s})}_{m,q}}} \right\}. \tag{5.184}$$

Führt man schließlich die rekursive Beobachtungsfunktion

$$f_{\mathrm{O},L_\mathrm{R}}^{(\mathrm{R})} : \mathbb{R}^{[2L_\mathrm{R}+3]Q} \to \mathbb{R}^{Q},\ f_{\mathrm{O},L_\mathrm{R}}^{(\mathrm{R})}\left(\mathbf{x}_{m:m-L_\mathrm{R}+1}^{(\mathrm{s})}, \bar{\mathbf{h}}_{0:L_\mathrm{R}-1}, \mathbf{y}_{m-L_\mathrm{R}}^{(\mathrm{s})}, \mathbf{n}_{m}^{(\mathrm{s})}, \mathbf{n}_{m-L_\mathrm{R}}^{(\mathrm{s})}\right) := \ln\left\{\sum_{m'=0}^{L_\mathrm{R}-1} e^{\mathbf{x}_{m-m'}^{(\mathrm{s})}+\bar{\mathbf{h}}_{m'}} + e^{-\frac{2L_\mathrm{R}B}{\tau_h}} \max\left[e^{\mathbf{y}_{m-L_\mathrm{R}}^{(\mathrm{s})}} - e^{\mathbf{n}_{m-L_\mathrm{R}}^{(\mathrm{s})}}, \mathbf{0}\right] + e^{\mathbf{n}_{m}^{(\mathrm{s})}}\right\} \tag{5.185}$$

ein, so ergibt sich ein rekursives Beobachtungsmodell in Vektornotation

$$\mathbf{y}_m^{(\mathrm{s})} = f_{\mathrm{O},L_\mathrm{R}}^{(\mathrm{R})}\left(\mathbf{x}_{m:m-L_\mathrm{R}+1}^{(\mathrm{s})}, \bar{\mathbf{h}}_{0:L_\mathrm{R}-1}, \mathbf{y}_{m-L_\mathrm{R}}^{(\mathrm{s})}, \mathbf{n}_{m}^{(\mathrm{s})}, \mathbf{n}_{m-L_\mathrm{R}}^{(\mathrm{s})}\right) + \mathbf{v}_{m,L_\mathrm{R}}^{(\mathrm{s,R})}. \tag{5.186}$$

Für den Fall $L_\mathrm{R} << L_H$ wird die Anzahl notwendiger Auswertungen der Exponentialfunktion gegenüber (5.130) deutlich reduziert.

In Abwesenheit von Hintergrundstörungen lässt sich auch die rekursive Beobachtungsfunktion durch Bildung des Grenzwertes von (5.185) für $\mathbf{n}_m^{(\mathrm{s})}, \mathbf{n}_{m-L_\mathrm{R}}^{(\mathrm{s})} \to (-\infty, \ldots, -\infty)^T$ zu

$$\tilde{f}_{\mathrm{O},L_\mathrm{R}}^{(\mathrm{R})} : \mathbb{R}^{[2L_\mathrm{R}+1]Q} \to \mathbb{R}^{Q},\quad \tilde{f}_{\mathrm{O},L_\mathrm{R}}^{(\mathrm{R})}\left(\mathbf{x}_{m:m-L_\mathrm{R}+1}^{(\mathrm{s})}, \bar{\mathbf{h}}_{0:L_\mathrm{R}-1}, \mathbf{y}_{m-L_\mathrm{R}}^{(\mathrm{s})}\right) := \ln\left\{\sum_{m'=0}^{L_\mathrm{R}-1} e^{\mathbf{x}_{m-m'}^{(\mathrm{s})}+\bar{\mathbf{h}}_{m'}} + e^{\left(-\frac{2L_\mathrm{R}B}{\tau_h}\cdot\mathbf{1}+\mathbf{y}_{m-L_\mathrm{R}}^{(\mathrm{s})}\right)}\right\} \tag{5.187}$$

vereinfachen, wobei $\mathbf{1} := (1, \ldots, 1)^T$. Die LMSK-Vektoren des verhallten Sprachsignals lassen sich dann durch

$$\mathbf{y}_m^{(\mathrm{s})} = \mathbf{s}_m^{(\mathrm{s})} \approx \tilde{f}_{\mathrm{O},L_\mathrm{R}}^{(\mathrm{R})}\left(\mathbf{x}_{m:m-L_\mathrm{R}+1}^{(\mathrm{s})}, \bar{\mathbf{h}}_{0:L_\mathrm{R}-1}, \mathbf{y}_{m-L_\mathrm{R}}^{(\mathrm{s})}\right). \tag{5.188}$$

annähern. Die qualitative Güte dieser Approximation wird bei dem Vergleich der Trajektorie der wahren LMSK-Vektoren eines beispielhaften Sprachsignals mit den entsprechenden Näherungen für $L_\mathrm{R} = 1$ bzw. $L_\mathrm{R} = 6$ in Abb. 5.5c bzw. Abb. 5.5d deutlich. Es lässt sich beobachten, dass die Approximation durch die rekursive Beobachtungsfunktion für $L_\mathrm{R} = 1$ im Vergleich zu der mit der nichtrekursive Beobachtungsfunktion (5.132) deutlich genauer ist und dass sehr feine Details nachgebildet werden können. Mit wachsenden Werten von L_R wird der Verlauf der Trajektorie immer glatter und nähert sich für $L_\mathrm{R} \to L_H$ dem in Abb. 5.5b an, da in dem Fall die rekursive annähernd in die nichtrekursive Beobachtungsfunktion übergeht.

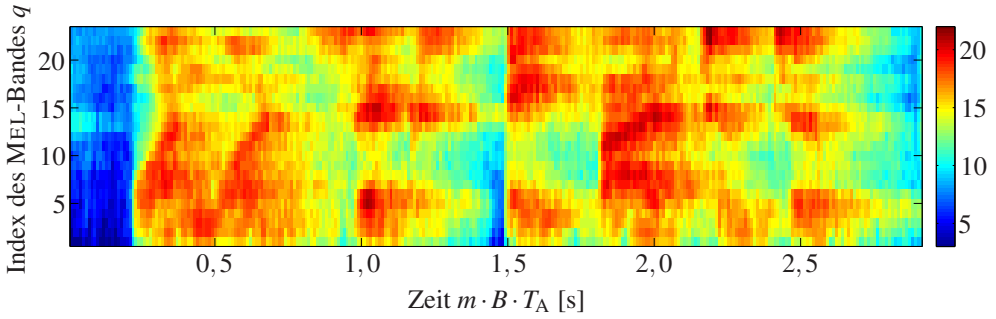
5.2.5. Modellierung des Beobachtungsfehlers

In diesem Abschnitt geht es um die Modellierung der beiden Beobachtungsfehler

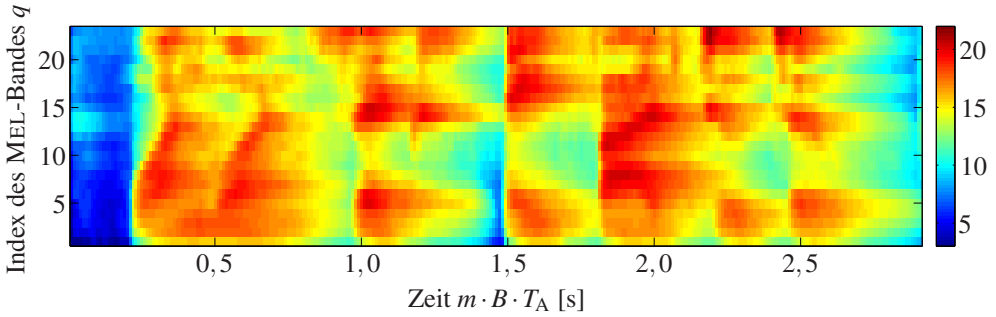
$$\mathbf{v}_m^{(\mathrm{s})} = \mathbf{y}_m^{(\mathrm{s})} - f_\mathrm{O}\left(\mathbf{x}_{m:m-L_H}^{(\mathrm{s})}, \bar{\mathbf{h}}_{0:L_H}, \mathbf{n}_m^{(\mathrm{s})}\right) \tag{5.189}$$

und

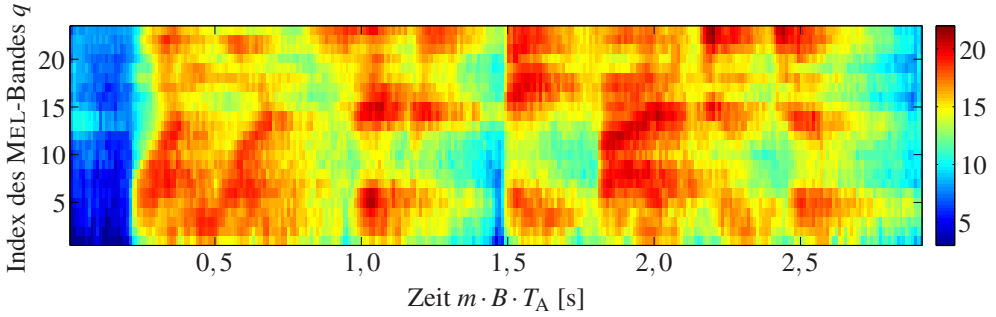
$$\mathbf{v}_{m,L_\mathrm{R}}^{(\mathrm{s,R})} = \mathbf{y}_m^{(\mathrm{s})} - f_{\mathrm{O},L_\mathrm{R}}^{(\mathrm{R})}\left(\mathbf{x}_{m:m-L_\mathrm{R}+1}^{(\mathrm{s})}, \bar{\mathbf{h}}_{0:L_\mathrm{R}-1}, \mathbf{y}_{m-L_\mathrm{R}}^{(\mathrm{s})}, \mathbf{n}_m^{(\mathrm{s})}, \mathbf{n}_{m-L_\mathrm{R}}^{(\mathrm{s})}\right), \tag{5.190}$$



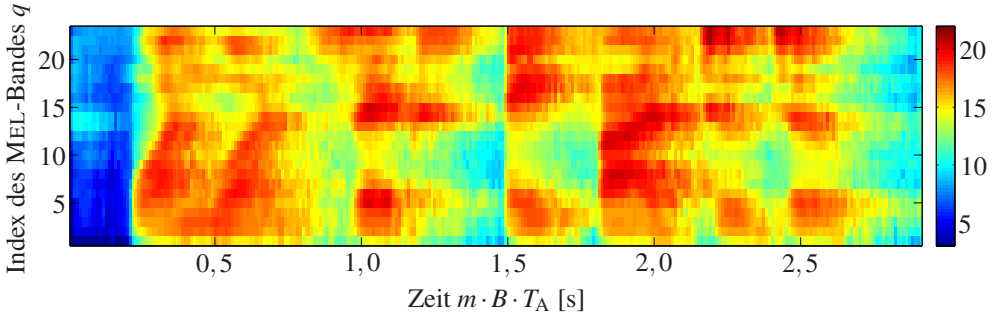
(a) Trajektorie der log-MEL-spektralen Merkmale $s^{(\mathrm{s})}_{m,q}$ des verhallten Sprachsignals



(b) Trajektorie der approximativen log-MEL-spektralen Merkmale $\hat{s}^{(\mathrm{s})}_{m,q}$ des verhallten Sprachsignals berechnet mit der nichtrekursiven Beobachtungsfunktion (5.132)



(c) Trajektorie der approximativen log-MEL-spektralen Merkmale $\hat{s}^{(\mathrm{s})}_{m,q}$ des verhallten Sprachsignals berechnet mit der rekursiven Beobachtungsfunktion ($L_{\mathrm{R}} = 1$)



(d) Trajektorie der approximativen log-MEL-spektralen Merkmale $\hat{s}^{(\mathrm{s})}_{m,q}$ des verhallten Sprachsignals berechnet mit der nichtrekursiven Beobachtungsfunktion ($L_{\mathrm{R}} = 6$)

Abbildung 5.5.: *Trajektorien der log-*MEL*-spektralen Merkmale eines beispielhaften verhallten Sprachsignals (selbes Signal wie in Abb. 2.5) und Approximationen durch unterschiedliche Beobachtungsmodelle.*

die in den Beobachtungsmodellen (5.131) und (5.186) auftreten. Dabei muss zunächst einmal berücksichtigt werden, dass bisher von der Annahme ausgegegangen wurde, dass die im Argument der Beobachtungsfunktionen auftretenden Terme $\bar{\mathbf{h}}_{0:L_H}$ bzw. $\bar{\mathbf{h}}_{0:L_\mathrm{R}-1}$ auf der Grundlage einer bekannten, zeitinvarianten Raumimpulsantwort berechnet werden. In der Praxis wird diese Annahme jedoch in der Regel nicht erfüllt sein, so dass die für die tatsächliche Auswertung der beiden Beobachtungsfunktionen f_O und $f_{\mathrm{O},L_\mathrm{R}}^{(\mathrm{R})}$ benötigten wahren, zeitvarianten Koeffizienten der RIA im log-MEL-spektralen Bereich $\bar{\mathbf{h}}_{0:L_H}$ nicht zur Verfügung stehen. Statt dessen werden diese Koeffizienten durch die auf dem RIA-Modell (5.141) und einer Schätzung der RIA-Parameter $\hat{\tau}_h$, $\hat{\sigma}_h$ und $\hat{L}_H$ basierenden Erwartungswerte

$$\hat{\boldsymbol{\mu}}_{\breve{\mathbf{h}}_{0:\hat{L}_H}} := \boldsymbol{\mu}_{\breve{\mathbf{h}}_{0:\hat{L}_H}}(\hat{\tau}_h, \hat{\sigma}_h) \tag{5.191}$$

ersetzt. Insofern sind für ein realistisches Szenario an Stelle der Fehler $\mathbf{v}_m^{(\mathrm{s})}$ und $\mathbf{v}_{m,L_\mathrm{R}}^{(\mathrm{s,R})}$ vielmehr die beiden Fehler

$$\hat{\mathbf{v}}_m^{(\mathrm{s})} := \mathbf{y}_m^{(\mathrm{s})} - f_\mathrm{O}\left(\mathbf{x}_{m:m-\hat{L}_H}^{(\mathrm{s})}, \hat{\boldsymbol{\mu}}_{\breve{\mathbf{h}}_{0:\hat{L}_H}}, \mathbf{n}_m^{(\mathrm{s})}\right) \tag{5.192}$$

und

$$\hat{\mathbf{v}}_{m,L_\mathrm{R}}^{(\mathrm{s,R})} := \mathbf{y}_m^{(\mathrm{s})} - f_{\mathrm{O},L_\mathrm{R}}^{(\mathrm{R})}\left(\mathbf{x}_{m:m-L_\mathrm{R}+1}^{(\mathrm{s})}, \hat{\boldsymbol{\mu}}_{\breve{\mathbf{h}}_{0:L_\mathrm{R}-1}}, \mathbf{y}_{m-L_\mathrm{R}}^{(\mathrm{s})}, \mathbf{n}_m^{(\mathrm{s})}, \mathbf{n}_{m-L_\mathrm{R}}^{(\mathrm{s})}\right) \tag{5.193}$$

interessant. Sie berücksichtigen sowohl Unzulänglichkeiten des RIA-Modells als auch Fehlschätzungen der Modellparameter. Da eine genaue analytische Beschreibung dieser Fehler sehr kompliziert ist, wird in dieser Arbeit ein stark vereinfachter, approximativer Ansatz verfolgt. Demnach werden beide Beobachtungsfehler als Realisierungen von stationären, weißen GAUSS'schen Zufallsprozessen gemäß

$$p\left(\hat{\mathbf{v}}_m^{(\mathrm{s})}\right) := \mathcal{N}\left(\hat{\mathbf{v}}_m^{(\mathrm{s})}; \boldsymbol{\mu}_{\hat{\mathbf{v}}^{(\mathrm{s})}}, \boldsymbol{\Sigma}_{\hat{\mathbf{v}}^{(\mathrm{s})}}\right) \tag{5.194}$$

$$p\left(\hat{\mathbf{v}}_{m,L_\mathrm{R}}^{(\mathrm{s,R})}\right) := \mathcal{N}\left(\hat{\mathbf{v}}_{m,L_\mathrm{R}}^{(\mathrm{s,R})}; \boldsymbol{\mu}_{\hat{\mathbf{v}}_{L_\mathrm{R}}^{(\mathrm{s,R})}}, \boldsymbol{\Sigma}_{\hat{\mathbf{v}}_{L_\mathrm{R}}^{(\mathrm{s,R})}}\right) \tag{5.195}$$

modelliert, was die Berechnung der Inferenz (siehe Kap. 5.3) ungemein vereinfacht.

Unter der weiteren Annahme der Ergodizität der Zufallsprozesse lassen sich die Parameter der Beobachtungsfehler $\boldsymbol{\mu}_{\hat{\mathbf{v}}^{(\mathrm{s})}}$ und $\boldsymbol{\Sigma}_{\hat{\mathbf{v}}^{(\mathrm{s})}}$ sowie $\boldsymbol{\mu}_{\hat{\mathbf{v}}_{L_\mathrm{R}}^{(\mathrm{s,R})}}$ und $\boldsymbol{\Sigma}_{\hat{\mathbf{v}}_{L_\mathrm{R}}^{(\mathrm{s,R})}}$ unter Verwendung von Stereotrainingsdaten, d.h. sauberen Sprachsignalen samt ihren verhallten und gestörten Versionen, vor der eigentlichen Merkmalsverbesserung empirisch berechnen. Um diese Schätzwerte sinnvoll verwenden zu können ist zu beachten, dass vor der Merkmalsverbesserung eine Normierung des Eingangssignals $y(l)$ stattfinden muss, so dass $\sigma_s^2 = \sigma_x^2$ näherungsweise gilt (siehe Kap. 5.2.3). Der Skalierungsfaktor für die RIA σ_h muss in dem Fall entsprechend (5.171) bestimmt werden.

Typischerweise sind die benötigten Stereotrainingsdaten, welche am Einsatzort des Sprachererkenners aufgenommen wurden, jedoch nicht vorhanden. Zumindest für den störungsfreien Fall, d.h. $n_{m,q}^{(\mathrm{s})} \ll x_{m,q}^{(\mathrm{s})} \;\; \forall m,q$, bietet sich die Möglichkeit, die erforderlichen Stereotrainingsdaten künstlich zu erzeugen. Dieses lässt sich beispielsweise bewerkstelligen, indem man

saubere Sprachsignale mit künstlichen RIAs faltet, welche mit der sogenannten Spiegelquellenmethode [All79] erzeugt werden. Bei dieser Methode wird die Schallausbreitung vom Sprecher zum Mikrophon unter der stark vereinfachten Annahme eines quaderförmigen, leeren Raumes mit starren Wänden simuliert. Die berechnete zeitinvariante RIA ist abhängig von der Position des Sprechers und des Mikrophons, der Raumgeometrie und den Absorptionseigenschaften der Wände. Zu gegebener Raumgeometrie sowie der Position des Sprechers und des Mikrophons lassen sich gemäß der Formel von SABINE [Kut00] die Absorbtionseigenschaften der Wände derart bestimmen, dass der simulierte Raum approximativ eine gewünschte Nachhallzeit aufweist. Um flexibel auf beliebige Einsatzorte des Erkenners vorbereitet zu sein, lassen sich auf diese Weise vorab Parameter des Beobachtungsmodells für eine relevante diskrete Menge von vorgegebenen Nachhallzeiten bzw. Abklingkonstanten berechnen, wobei jeweils zur Berechnung der Koeffizienten der Raumimpulsantwort $\hat{\boldsymbol{\mu}}_{\check{\mathbf{h}}_{0:\hat{L}_H}}$ in (5.192) und (5.193) der Skalierungsfaktor $\hat{\sigma}_h$ gemäß (5.171) bestimmt wird. Während der Merkmalsverbesserung können dann, beruhend auf einer Schätzung der Nachhallzeit, die am besten passenden Parameter ausgewählt werden. Zur Berücksichtigung möglichst vieler unterschiedlicher Erkennungsszenarien wird hier vorgeschlagen, viele unterschiedliche RIAs zur Erzeugung der Stereotrainingsdaten zu verwenden, die sich in der Anordnung des Sprechers und Mikrophons im Raum unterscheiden. Es ist weiterhin sinnvoll die Nachhallzeit innerhalb eines gewissen Intervalls um den vorgegebenen Wert zufällig zu variieren, um während der Merkmalsverbesserung auftretende Schätzfehler der Nachhallzeit in Betracht zu ziehen.

Experimentelle Untersuchungen zur Validierung der gemachten Annahmen (5.194) und (5.195) für den störungsfreien Fall auf ausgewählten Sprachdatenbanken folgen in Kap. 6.4.

Für den Fall, dass neben dem Nachhall zusätzlich Hintergrundstörungen in dem Mikrophonsignal vorhanden sind, ist die Modellierung des Beobachtungsfehlers mit Hilfe eines GAUSS'schen Zufallsprozesses gemäß (5.194) für das nichtrekursive bzw. gemäß (5.195) für das rekursive Beobachtungsmodell eigentlich nicht mehr sinnvoll. Der Beobachtungsfehler $v_{m,q}^{(\mathrm{s})}$ bzw. $v_{m,q,L_\mathrm{R}}^{(\mathrm{s,R})}$ ist dann in hohem Maße abhängig vom lokalen Signal-zu-Rauschleistungsverhältnis (engl. ***S**ignal-to-**N**oise **R**atio* (*SNR*)) zum Zeitpunkt m im q-ten MEL-Band, wobei grob drei Fälle zu unterscheiden sind.

Ist das lokale *SNR* sehr niedrig, dann dominiert die Störung stark im Verhältnis zum Sprachanteil, so dass für den MEL-spektralen Koeffizienten $\mathscr{Y}_{m,q}$ in sehr guter Näherung

$$\mathscr{Y}_{m,q} \approx \mathscr{N}_{m,q} \tag{5.196}$$

$$\mathscr{N}_{m,q} \gg C_\mathrm{E} \cdot \bar{\mathscr{H}}_{m',q} \mathscr{X}_{m-m',q} \quad \text{für } m' \in \{0,\ldots,L_H\} \tag{5.197}$$

gilt. Aufgrund dessen verschwindet der in (5.125) definierte Approximationsfehler $\mathscr{E}_{m,q}$ näherungsweise, so dass der resultierende Beobachtungsfehler $v_{m,q}^{(\mathrm{s})}$ gemäß (5.129) relativ klein ist. Eine ähnliche Argumentation lässt sich für den in (5.182) definierten Approximationsfehler $\mathscr{E}_{m,L_\mathrm{R},q}^{(\mathrm{R})}$ und den resultierenden Beobachtungsfehler $v_{m,L_\mathrm{R},q}^{(\mathrm{s,R})}$ im Falle der Verwendung der rekursiven Beobachtungsfunktion führen.

Ist im Gegensatz dazu die Sprache dominant, so liegt eine ähnliche Situation wie im störungsfreien Fall vor. Der Beobachtungsfehler ist dann im Vergleich zum Fall zuvor im Mittel deutlich größer. Das liegt zum einen daran, dass der Approximationsfehler in (5.117) bedingt durch die Vernachlässigung der Kreuzterme im ersten Summenterm von (5.116) relativ groß

ist. Zum anderen ist die Approximation in (5.124) durch die Ersetzung der in (5.121) auftretende Terme $\left|h_{k,k}(m')\right|^2$ durch ihre Mittelwerte über das q-te MEL-Band $\tilde{\mathscr{H}}_{m',q}$ relativ grob.

In dem Fall, dass das Sprach- und das Störsignal lokal eine annähernd gleiche Leistung aufweisen, ist der mittlere Beobachtungsfehler im Allgemeinen am größten. Denn dann wirkt sich zusätzlich die Vernachlässigung des zweiten Summenterms in (5.116) im Hinblick auf die Approximation (5.117) auf den Approximationsfehler aus.

Unter Berücksichtigung dieser Tatsachen hängt ein mit Hilfe von Trainingsdaten empirisch bestimmtes Histogramm des Beobachtungsfehlers $\hat{v}_{m,q}^{(\mathrm{s})}$ bzw. $\hat{v}_{m,L_\mathrm{R},q}^{(\mathrm{s,R})}$ nicht nur in hohem Maße von der Art der Störung und dem *SNR* ab, sondern auch vom Anteil der Sprachpausen in den Trainingsäußerungen. Mit abnehmendem *SNR* und zunehmendem Anteil der Sprachpausen wird das Histogramm immer steilgipfliger, sodass es nicht mehr hinreichend genau durch eine GAUSS-Verteilungsdichtefunktion approximiert werden kann.

Im Bewusstsein dessen, dass diese Art der Lösung sehr unzufriedenstellend und bei weitem nicht optimal ist, wird in dieser Arbeit der stark vereinfachte Ansatz verfolgt, bei Vorhandensein der Störung dieselben Parameter des Beobachtungsfehlers wie im Fall ohne Störung zu nutzen. Er ist zumindest für sehr hohe Werte des *SNR* gerechtfertigt. Eine Entwicklung genauerer Modelle für den Beobachtungsfehler zur Berücksichtigung des Einflusses der Störung bleibt Gegenstand zukünftiger Forschung.

5.3. Inferenz

Nachdem zu Beginn von Kap. 5 das Konzept der BAYES'schen Merkmalsverbesserung vorgestellt und in Kap. 5.1 und Kap. 5.2 jeweils das dazu verwendete A-priori-Modell und Beobachtungsmodell ausführlich beschrieben wurde, widmet sich dieser Abschnitt nun der praktischen Umsetzung der Merkmalsverbesserung.

Zur Erinnerung sei noch einmal darauf hingewiesen, dass der Kern der BAYES'schen Merkmalsverbesserung durch die rekursive Bestimmung der A-posteriori-Verteilungsdichtefunktion $p\left(\mathbf{z}_m^{(\mathrm{s})}\middle|\mathbf{y}_{1:m}^{(\mathrm{s})}\right)$ gegeben ist. Im Allgemeinen gestaltet sich die dazu erforderliche rekursive Berechnung der Prädiktion und Aktualisierung gemäß der beiden Gleichungen (5.6) und (5.7) sehr schwierig, da für den Fall einer beliebigen Form der Verteilungsdichtefunktion $p\left(\mathbf{z}_{m-1}^{(\mathrm{s})}\middle|\mathbf{y}_{1:m-1}^{(\mathrm{s})}\right)$ keine vernünftig handhabbare analytische Lösung für $p\left(\mathbf{z}_m^{(\mathrm{s})}\middle|\mathbf{y}_{1:m-1}^{(\mathrm{s})}\right)$ und $p\left(\mathbf{z}_m^{(\mathrm{s})}\middle|\mathbf{y}_{1:m}^{(\mathrm{s})}\right)$ angegeben werden kann.

Eine Möglichkeit zur Lösung des Problems besteht dann in der Anwendung von MONTE-CARLO-Methoden zur approximativen Berechnung der gesuchten Verteilungsdichtefunktionen. Eine ausführliche und anschauliche Beschreibung solcher Verfahren findet sich beispielsweise in [AMGC02]. Ihre Idee basiert auf der approximativen Darstellung einer Verteilungsdichtefunktion mit Hilfe einer Menge von gewichteten Stichproben, sogenannten Partikeln, welchen dieselbe Verteilungsdichtefunktion zugrunde liegt. Ein entscheidender Nachteil liegt jedoch in der Tatsache, dass die Anzahl der benötigten Partikel, und damit auch der Rechenaufwand, für eine hinreichend genaue Approximation einer Verteilungsdichtefunktion im Allgemeinen exponentiell mit der Dimension der Zufallsvektoren wächst. Da die

Dimension des hier betrachteten Merkmalsvektors

$$\mathbf{z}_m^{(\mathrm{s})} = \left[\left(\mathbf{x}_m^{(\mathrm{s})} \right)^T, \ldots, \left(\mathbf{x}_{m-L_\mathrm{C}+1}^{(\mathrm{s})} \right)^T, \left(\mathbf{n}_m^{(\mathrm{s})} \right)^T \right]^T \tag{5.198}$$

durch $(L_\mathrm{C}+1)\,Q$ gegeben ist, wobei gemäß Tab. 2.1 $Q = 23$ gilt, werden derartige Verfahren hier nicht weiter betrachtet.

Hingegen wird hier ein anderer vereinfachter, approximativer Ansatz verfolgt, dessen Motivation im Folgenden schrittweise verdeutlicht wird.

5.3.1. Iteratives erweitertes KALMAN-Filter

Geht man vorläufig von der approximativen Annahme aus, dass die A-priori-Verteilungsdichtefunktion $p\left(\mathbf{z}_{m-1}^{(\mathrm{s})} \middle| \mathbf{y}_{1:m-1}^{(\mathrm{s})} \right)$ durch eine GAUSS-Verteilungsdichtefunktion gemäß

$$p\left(\mathbf{z}_{m-1}^{(\mathrm{s})} \middle| \mathbf{y}_{1:m-1}^{(\mathrm{s})} \right) = \mathcal{N}\left(\mathbf{z}_{m-1}^{(\mathrm{s})}; \hat{\mathbf{z}}_{m-1|m-1}^{(\mathrm{s})}, \hat{\boldsymbol{\Sigma}}_{\mathbf{z}_{m-1|m-1}^{(\mathrm{s})}} \right) \tag{5.199}$$

gegeben ist, so lässt sich zeigen, dass für den Fall eines linearen A-priori-Modells und Beobachtungsmodells sowie GAUSS-verteilten Prädiktions- und Beobachtungsfehlern die a posteriori-Verteilungsdichtefunktion $p\left(\mathbf{z}_m^{(\mathrm{s})} \middle| \mathbf{y}_{1:m}^{(\mathrm{s})} \right)$ selbst wieder eine GAUSS-Verteilung darstellt, deren Mittelwert und Kovarianzmatrix mit Hilfe eines KALMAN-Filters berechnet werden können [BSLK01]. In einer solchen Situation reduziert sich die Inferenz auf die Berechnung der ersten beiden zentralen Momente.

In dem hier betrachteten Fall sind die dazu benötigten Voraussetzungen insofern nicht erfüllt, als dass das A-priori-Modell zwar aus linearen Teilmodellen besteht, als Ganzes aber nichtlinear ist. Zudem sind beide alternativen Beobachtungsfunktionen f_O und $f_{\mathrm{O},L_\mathrm{R}}^{(\mathrm{R})}$ nichtlinear. Eine approximative Lösung für die beiden ersten zentralen Momente $\hat{\mathbf{z}}_{m|m,i}^{(\mathrm{s})}$ und $\hat{\boldsymbol{\Sigma}}_{\mathbf{z}_{m|m,i}^{(\mathrm{s})}}$ der auf das i-te Teilmodell bedingten A-posteriori-Verteilungsdichtefunktion

$$p\left(\mathbf{z}_m^{(\mathrm{s})} \middle| \mathbf{y}_{1:m}^{(\mathrm{s})}, \zeta_m = i \right) \approx \mathcal{N}\left(\mathbf{z}_m^{(\mathrm{s})}; \hat{\mathbf{z}}_{m|m,i}^{(\mathrm{s})}, \hat{\boldsymbol{\Sigma}}_{\mathbf{z}_{m|m,i}^{(\mathrm{s})}} \right) \tag{5.200}$$

lässt sich dann mit einem sogenannten iterativen erweiterten KALMAN-Filter (engl. ***I**terated **E**xtended* **K**ALMAN ***F**ilter* (*IEKF*)) [BSLK01] gemäß Alg. 4 berechnen. Als Eingabe werden zusätzlich zu den bereits angesprochenen beiden zentralen Momenten der A-priori-Verteilungsdichtefunktion $\hat{\mathbf{z}}_{m-1|m-1}^{(\mathrm{s})}$ und $\hat{\boldsymbol{\Sigma}}_{\mathbf{z}_{m-1|m-1}^{(\mathrm{s})}}$ unter anderem die Schätzwerte für die Mittelwertvektoren und Kovarianzmatrizen des sauberen Sprachsignals und des Störsignals vergangener Zeitpunkte benötigt. Es ist wichtig zu bemerken, dass diese Schätzungen in den vorhergehenden Inferenzschritten berechnet und zwischengespeichert werden müssen. Weiterhin hängt es von der verwendeten Beobachtungsfunktion ab, welche dieser Schätzungen tatsächlich benötigt werden.

Im *IEKF* wird zunächst abhängig von dem Segmentindex m und dem Teilmodellindex i die Prädiktion basierend auf (5.13) durchgeführt. Dieser Schritt ist aufgrund der Linearität des A-priori-Teilmodells noch völlig identisch mit dem eines gewöhnlichen KALMAN-Filters.

Algorithmus 4: Iteratives erweitertes KALMAN-Filter

Eingabe: $\hat{\mathbf{z}}^{(\mathrm{s})}_{m-1|m-1}$, $\hat{\boldsymbol{\Sigma}}_{\mathbf{z}^{(\mathrm{s})}_{m-1|m-1}}$, $\hat{\mathbf{x}}^{(\mathrm{s})}_{m-\hat{L}_H:m-L_\mathrm{C}+1}$, $\hat{\boldsymbol{\Sigma}}_{\breve{\mathbf{x}}^{(\mathrm{s})}_{m-\hat{L}_H:m-L_\mathrm{C}+1}}$,
$\hat{\mathbf{n}}^{(\mathrm{s})}_{m-L_\mathrm{C}}$, $\hat{\boldsymbol{\Sigma}}_{\breve{\mathbf{n}}^{(\mathrm{s})}_{m-L_\mathrm{C}}}$, $\hat{\mathbf{x}}^{(\mathrm{s})}_{-L_\mathrm{C}+2:0}$, $\hat{\boldsymbol{\Sigma}}_{\breve{\mathbf{x}}^{(\mathrm{s})}_{-L_\mathrm{C}+2:0}}$, $\mathbf{y}^{(\mathrm{s})}_{m-L_\mathrm{C}}$, $\hat{\boldsymbol{\mu}}_{\breve{\mathbf{h}}_{0:\hat{L}_H}}$, m, k.

Ausgabe: $\hat{\mathbf{z}}^{(\mathrm{s})}_{m|m,i}$, $\hat{\boldsymbol{\Sigma}}_{\mathbf{z}^{(\mathrm{s})}_{m|m,i}}$, $\hat{\mathbf{y}}^{(\mathrm{s}),[1]}_{m,i}$, $\hat{\boldsymbol{\Sigma}}_{\mathbf{y}^{(\mathrm{s}),[1]}_{m,i}}$.

1. **Prädiktion:**

   Wenn $m \leq L_\mathrm{AR}$ **dann**

   - Initialisiere den Mittelwertvektor $\hat{\mathbf{z}}^{(\mathrm{s})}_{m|m-1,i}$ und die Kovarianzmatrix $\hat{\boldsymbol{\Sigma}}_{\mathbf{z}^{(\mathrm{s})}_{m|m-1,i}}$ der prädiktiven Verteilungsdichtefunktion $p\left(\mathbf{z}^{(\mathrm{s})}_m \middle| \zeta_m = i\right)$ gemäß

$$\hat{\mathbf{z}}^{(\mathrm{s})}_{m|m-1,i} = \Big[\underbrace{(\boldsymbol{\mu}_{\mathbf{x},i})^T \ \ldots \ (\boldsymbol{\mu}_{\mathbf{x},i})^T}_{m\text{-mal}} \ \left(\hat{\mathbf{x}}^{(\mathrm{s})}_0\right)^T \ \ldots \ \left(\hat{\mathbf{x}}^{(\mathrm{s})}_{-L_\mathrm{C}+m+1}\right)^T \ (\boldsymbol{\mu}_\mathbf{n})^T\Big]^T \tag{5.201}$$

$$\hat{\boldsymbol{\Sigma}}_{\mathbf{z}^{(\mathrm{s})}_{m|m-1,i}} = \begin{bmatrix} \mathrm{blockdiag}\Big\{\overbrace{\boldsymbol{\Sigma}_{\mathbf{x},i}, \ldots, \boldsymbol{\Sigma}_{\mathbf{x},i}}^{m\text{-mal}}\Big\} & \mathbf{0} & \ldots & \ldots & \mathbf{0} \\ \mathbf{0} & \hat{\boldsymbol{\Sigma}}_{\breve{\mathbf{x}}^{(\mathrm{s})}_0} & \mathbf{0} & \ldots & \mathbf{0} \\ \vdots & \mathbf{0} & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \hat{\boldsymbol{\Sigma}}_{\breve{\mathbf{x}}^{(\mathrm{s})}_{-L_\mathrm{C}+m+1}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0} & \boldsymbol{\Sigma}_\mathbf{n} \end{bmatrix}. \tag{5.202}$$

   sonst

   - Berechne den Mittelwertvektor $\hat{\mathbf{z}}^{(\mathrm{s})}_{m|m-1,i}$ und die Kovarianzmatrix $\hat{\boldsymbol{\Sigma}}_{\mathbf{z}^{(\mathrm{s})}_{m|m-1,i}}$ der prädiktiven Verteilungsdichtefunktion $p\left(\mathbf{z}^{(\mathrm{s})}_m \middle| \mathbf{y}^{(\mathrm{s})}_{1:m-1}, \zeta_m = i\right)$ gemäß

$$\hat{\mathbf{z}}^{(\mathrm{s})}_{m|m-1,i} = \mathbf{A}_{\mathbf{z},i}\hat{\mathbf{z}}^{(\mathrm{s})}_{m-1|m-1} + \mathbf{b}_{\mathbf{z},i} \tag{5.203}$$

$$\hat{\boldsymbol{\Sigma}}_{\mathbf{z}^{(\mathrm{s})}_{m|m-1,i}} = \mathbf{A}_{\mathbf{z},i}\hat{\boldsymbol{\Sigma}}_{\mathbf{z}^{(\mathrm{s})}_{m-1|m-1}}(\mathbf{A}_{\mathbf{z},i})^T + \mathbf{V}_{\mathbf{z},i} \tag{5.204}$$

   mit

$$\mathbf{A}_{\mathbf{z},i} := \begin{bmatrix} \mathbf{A}_{i,1} & \ldots & \mathbf{A}_{i,L_\mathrm{AR}} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{I} & \mathbf{0} & \ldots & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \ddots & \mathbf{0} & \mathbf{0} & \vdots \\ \vdots & \mathbf{0} & \ddots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \ddots & \mathbf{I} & \mathbf{0} & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0} & \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \ldots & \ldots & \ldots & \mathbf{0} & \mathbf{0} \end{bmatrix}, \quad \mathbf{V}_{\mathbf{z},i} := \begin{bmatrix} \mathbf{V}_i & \mathbf{0} & \ldots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0} & \boldsymbol{\Sigma}_\mathbf{n} \end{bmatrix}, \tag{5.205}$$

$$\mathbf{b}_{\mathbf{z},i} := \left[(\mathbf{b}_i)^T \ \mathbf{0} \ \ldots \ \mathbf{0} \ (\boldsymbol{\mu}_\mathbf{n})^T\right]^T. \tag{5.206}$$

   Ende wenn

2. **Aktualisierung:**

   a) Initialisiere die Linearisierungsstelle der Beobachtungsfunktion mit dem Prädiktionsvektor gemäß

   $$\hat{\mathbf{z}}_{m|m,i}^{(\mathrm{s}),[1]} = \hat{\mathbf{z}}_{m|m-1,i}^{(\mathrm{s})}. \tag{5.207}$$

   b) Iteriere die Linearisierungsstelle:

   Für $r = 1..R$

   i. Berechne die prädizierte Beobachtung $\hat{\mathbf{y}}_{m,i}^{(\mathrm{s}),[r]}$, die zugehörige Kovarianzmatrix $\hat{\boldsymbol{\Sigma}}_{\mathbf{y}_{m,i}^{(\mathrm{s}),[r]}}$ sowie die JACOBI-Matrix $\mathbf{H}_{\hat{\mathbf{z}}_{m|m,i}^{(\mathrm{s}),[r]}}$:

   Wenn nichtrekursive Beobachtungsfunktion (5.130) verwendet wird **dann**

   - Verwende Alg. 5:

     Eingabe: $\hat{\mathbf{z}}_{m|m,i}^{(\mathrm{s}),[r]}$, $\hat{\mathbf{x}}_{m-\hat{L}_H:m-L_\mathrm{C}}^{(\mathrm{s})}$, $\hat{\boldsymbol{\Sigma}}_{\check{\mathbf{x}}_{m-\hat{L}_H:m-L_\mathrm{C}}^{(\mathrm{s})}}$, $\hat{\boldsymbol{\mu}}_{\check{\mathbf{h}}_{0:\hat{L}_H}}$.

     Ausgabe: $\hat{\mathbf{y}}_{m,i}^{(\mathrm{s}),[r]}$, $\hat{\boldsymbol{\Sigma}}_{\mathbf{y}_{m,i}^{(\mathrm{s}),[r]}}$, $\mathbf{H}_{\hat{\mathbf{z}}_{m|m,i}^{(\mathrm{s}),[r]}}$.

   Sonst (d.h. wenn rekursive Beobachtungsfunktion (5.185) verwendet wird)

   - Verwende Alg. 6:

     Eingabe: $\hat{\mathbf{z}}_{m|m,i}^{(\mathrm{s}),[r]}$, $\mathbf{y}_{m-L_\mathrm{C}}^{(\mathrm{s})}$, $\hat{\mathbf{n}}_{m-L_\mathrm{C}}^{(\mathrm{s})}$, $\hat{\boldsymbol{\Sigma}}_{\check{\mathbf{n}}_{m-L_\mathrm{C}}^{(\mathrm{s})}}$, $\hat{\boldsymbol{\mu}}_{\check{\mathbf{h}}_{0:L_\mathrm{C}-1}}$.

     Ausgabe: $\hat{\mathbf{y}}_{m,i}^{(\mathrm{s}),[r]}$, $\hat{\boldsymbol{\Sigma}}_{\mathbf{y}_{m,i}^{(\mathrm{s}),[r]}}$, $\mathbf{H}_{\hat{\mathbf{z}}_{m|m,i}^{(\mathrm{s}),[r]}}$.

   Ende wenn

   ii. Aktualisiere die Linearisierungsstelle gemäß

   $$\hat{\mathbf{z}}_{m|m,i}^{(\mathrm{s}),[r+1]} = \hat{\mathbf{z}}_{m|m-1,i}^{(\mathrm{s})} + \mathbf{K}_{m,i}^{[r]} \left[\mathbf{y}_m^{(\mathrm{s})} - \hat{\mathbf{y}}_{m,i}^{(\mathrm{s}),[r]} + \mathbf{H}_{\hat{\mathbf{z}}_{m|m,i}^{(\mathrm{s}),[r]}} \left(\hat{\mathbf{z}}_{m|m,i}^{(\mathrm{s}),[r]} - \hat{\mathbf{z}}_{m|m-1,i}^{(\mathrm{s})} \right) \right] \tag{5.208}$$

   mit der KALMAN-Verstärkungsmatrix

   $$\mathbf{K}_{m,i}^{[r]} := \hat{\boldsymbol{\Sigma}}_{\mathbf{z}_{m|m-1,i}^{(\mathrm{s})}} \left(\mathbf{H}_{\hat{\mathbf{z}}_{m|m,i}^{(\mathrm{s}),[r]}} \right)^T \left(\hat{\boldsymbol{\Sigma}}_{\mathbf{y}_{m,i}^{(\mathrm{s}),[r]}} \right)^{-1}. \tag{5.209}$$

   Ende für

   c) Berechne den Mittelwertvektor $\hat{\mathbf{z}}_{m|m,i}^{(\mathrm{s})}$ und die Kovarianzmatrix $\hat{\boldsymbol{\Sigma}}_{\mathbf{z}_{m|m,i}^{(\mathrm{s})}}$ der A-posteriori-Verteilungsdichtefunktion $p\left(\mathbf{z}_m^{(\mathrm{s})} \middle| \mathbf{y}_{1:m}^{(\mathrm{s})}, \zeta_m = i \right)$ gemäß

   $$\hat{\mathbf{z}}_{m|m,i}^{(\mathrm{s})} = \hat{\mathbf{z}}_{m|m,i}^{(\mathrm{s}),[R+1]}, \tag{5.210}$$

   $$\hat{\boldsymbol{\Sigma}}_{\mathbf{z}_{m|m,i}^{(\mathrm{s})}} = \left(\mathbf{I} - \mathbf{K}_{m,i}^{[R]} \; \mathbf{H}_{\hat{\mathbf{z}}_{m|m,i}^{(\mathrm{s}),[R]}} \right) \hat{\boldsymbol{\Sigma}}_{\mathbf{z}_{m|m-1,i}^{(\mathrm{s})}}. \tag{5.211}$$

Der Grundgedanke besteht dabei darin, dass eine Zufallsvariable, die durch eine lineare Transformation aus einer normalverteilten Zufallsvariablen hervorgeht, selbst wieder normalverteilt ist. Bei der Berechnung der beiden Ausdrücke $\hat{\mathbf{z}}^{(\mathrm{s})}_{m|m-1,i}$ und $\hat{\boldsymbol{\Sigma}}_{\mathbf{z}^{(\mathrm{s})}_{m|m-1,i}}$ in (5.201) und (5.202) in Alg. 4 ist darauf zu achten, dass die beiden Terme

$$\left(\hat{\mathbf{x}}^{(\mathrm{s})}_{0}\right)^T \quad \ldots \quad \left(\hat{\mathbf{x}}^{(\mathrm{s})}_{-L_\mathrm{C}+m+1}\right)^T \tag{5.212}$$

$$\mathrm{blockdiag}\left\{\hat{\boldsymbol{\Sigma}}_{\breve{\mathbf{x}}^{(\mathrm{s})}_{0}}, \ldots, \hat{\boldsymbol{\Sigma}}_{\breve{\mathbf{x}}^{(\mathrm{s})}_{-L_\mathrm{C}+m+1}}\right\} \tag{5.213}$$

für $-L_\mathrm{C}+m+1>0$ zu ignorieren sind.

Zur Aktualisierung, dem zweiten Teilschritt, wird zur Ausnutzung dieses Prinzips die nichtlineare Beobachtungsfunktion anfangs an der Prädiktionsstelle linearisiert. Die Linearisierungstelle wird anschließend R Mal mit dem Ziel der Maximierung der A-posteriori-Verteilungsdichtefunktion $p\left(\mathbf{z}^{(\mathrm{s})}_{m}\middle|\mathbf{y}^{(\mathrm{s})}_{1:m}, \zeta_m = i\right)$ iterativ verbessert [BSLK01]. Dazu werden die ersten beiden zentralen Momente der prädiktiven Verteilungsdichtefunktion der Beobachtung $\mathbf{y}^{(\mathrm{s})}_{m}$ bedingt auf die Linearisierungsstelle bestimmt. Bei der Verwendung des nichtrekursiven Beobachtungsmodells geschieht dieses mit Hilfe von Alg. 5. Für das rekursive Beobachtungsmodell wird Alg. 6 herangezogen. Dabei ist zu beachten, dass die Rekursionslänge L_R gleich der Anzahl L_C der Merkmalsvektoren des sauberen Sprachsignals innerhalb des Zustandsvektors $\mathbf{z}^{(\mathrm{s})}_{m}$ gewählt wird.

Außerdem wird durch eine Betrachtung von Alg. 5 ersichtlich, dass beim nichtrekursiven Beobachtungsmodell die prädiktive Verteilungsdichtefunktion der Beobachtung $\mathbf{y}^{(\mathrm{s})}_{m}$ unter anderem mit Hilfe der geschätzten Merkmalsvektorfolge $\hat{\mathbf{x}}^{(\mathrm{s})}_{m-\hat{L}_H:m-L_\mathrm{C}}$ des sauberen Sprachsignals sowie der zugehörigen geschätzten Kovarianzmatrizen $\hat{\boldsymbol{\Sigma}}_{\breve{\mathbf{x}}^{(\mathrm{s})}_{m-\hat{L}_H:m-L_\mathrm{C}}}$ berechnet wird. Nimmt man beispielsweise eine Nachhallzeit T_{60} von 0,45 s an und approximiert die Länge $\hat{L}_h$ einer zugehörigen RIA gemäß (5.159) mit $\varepsilon_h = 10^{-3}$, so ergibt sich nach (5.112) für $\hat{L}_H$ ein Wert von 24. Bei der Bestimmung der Kovarianzmatrix $\hat{\boldsymbol{\Sigma}}_{\mathbf{y}^{(\mathrm{s}),[r]}_{m,i}}$ in (5.215) wird vereinfacht angenommen, dass die einzelnen Schätzvektoren der Sequenz $\hat{\mathbf{x}}^{(\mathrm{s})}_{m-\hat{L}_H:m-L_\mathrm{C}}$ sowohl untereinander als auch mit $\hat{\mathbf{z}}^{(\mathrm{s}),[r]}_{m|m,i}$ unkorreliert sind.

Hingegen werden bei der Verwendung des rekursiven Beobachtungsmodells in Alg. 6 statt der Schätzvektorfolge $\hat{\mathbf{x}}^{(\mathrm{s})}_{m-\hat{L}_H:m-L_\mathrm{C}}$ lediglich die vergangene Beobachtung $\mathbf{y}^{(\mathrm{s})}_{m-L_\mathrm{C}}$, der Schätzvektor $\hat{\mathbf{n}}^{(\mathrm{s})}_{m-L_\mathrm{C}}$ des zeitlich zurückliegenden Merkmalsvektors des Störsignals sowie die zugehörige Kovarianzmatrix $\hat{\boldsymbol{\Sigma}}_{\breve{\mathbf{n}}^{(\mathrm{s})}_{m-L_\mathrm{C}}}$ benötigt. Dabei soll noch einmal betont werden, dass damit im Vergleich zur Verwendung der nichtrekursiven Beobachtungsfunktion in Alg. 5 eine Reduktion des benötigten Rechen- und Speicheraufwands erzielt wird. Zur Berechnung der Kovarianzmatrix $\hat{\boldsymbol{\Sigma}}_{\mathbf{y}^{(\mathrm{s}),[r]}_{m,i}}$ in (5.221) wird nur noch angenommen, dass die beiden Schätzvektoren $\hat{\mathbf{n}}^{(\mathrm{s})}_{m-L_\mathrm{C}}$ und $\hat{\mathbf{z}}^{(\mathrm{s}),[r]}_{m|m,i}$ unkorreliert sind.

Algorithmus 5 Berechnung der ersten beiden zentralen Momente der Verteilungsdichtefunktion $p\left(\mathbf{y}_m^{(\mathrm{s})} \middle| \hat{\mathbf{z}}_{m|m,i}^{(\mathrm{s}),[r]}, \hat{\mathbf{x}}_{m-\hat{L}_H:m-L_\mathrm{C}}^{(\mathrm{s})}, \hat{\boldsymbol{\Sigma}}_{\breve{\mathbf{x}}_{m-\hat{L}_H:m-L_\mathrm{C}}^{(\mathrm{s})}}, \hat{\boldsymbol{\mu}}_{\breve{\mathbf{h}}_{0:\hat{L}_H}}\right)$ basierend auf der nicht rekursiven Beobachtungsfunktion (5.130).

Eingabe: $\hat{\mathbf{z}}_{m|m,i}^{(\mathrm{s}),[r]}, \hat{\mathbf{x}}_{m-\hat{L}_H:m-L_\mathrm{C}}^{(\mathrm{s})}, \hat{\boldsymbol{\Sigma}}_{\breve{\mathbf{x}}_{m-\hat{L}_H:m-L_\mathrm{C}}^{(\mathrm{s})}}, \hat{\boldsymbol{\mu}}_{\breve{\mathbf{h}}_{0:\hat{L}_H}}$.

Ausgabe: $\hat{\mathbf{y}}_{m,i}^{(\mathrm{s}),[r]}, \hat{\boldsymbol{\Sigma}}_{\mathbf{y}_{m,i}^{(\mathrm{s}),[r]}}, \mathbf{H}_{f_\mathrm{O},\hat{\mathbf{z}}_{m|m,i}^{(\mathrm{s}),[r]}}$.

- Berechne die prädizierte Beobachtung $\hat{\mathbf{y}}_{m,i}^{(\mathrm{s}),[r]}$ und die zugehörige Kovarianzmatrix $\hat{\boldsymbol{\Sigma}}_{\mathbf{y}_{m,i}^{(\mathrm{s}),[r]}}$ basierend auf der Linearisierungsstelle $\hat{\mathbf{z}}_{m|m,i}^{(\mathrm{s}),[r]}$ gemäß

$$\hat{\mathbf{y}}_{m,i}^{(\mathrm{s}),[r]} = f_\mathrm{O}\left(\hat{\boldsymbol{\chi}}_{m|m,i}^{(\mathrm{s}),[r]}, \hat{\mathbf{x}}_{m-L_\mathrm{C}:m-\hat{L}_H}^{(\mathrm{s})}, \hat{\boldsymbol{\mu}}_{\breve{\mathbf{h}}_{0:\hat{L}_H}}, \hat{\mathbf{n}}_{m|m,i}^{(\mathrm{s}),[r]}\right) + \boldsymbol{\mu}_{\hat{\mathbf{v}}^{(\mathrm{s})}} \tag{5.214}$$

$$\begin{aligned}\hat{\boldsymbol{\Sigma}}_{\mathbf{y}_{m,i}^{(\mathrm{s}),[r]}} = {} & \mathbf{H}_{f_\mathrm{O},\hat{\mathbf{z}}_{m|m,i}^{(\mathrm{s}),[r]}} \hat{\boldsymbol{\Sigma}}_{\mathbf{z}_{m|m-1,i}^{(\mathrm{s})}} \left(\mathbf{H}_{f_\mathrm{O},\hat{\mathbf{z}}_{m|m,i}^{(\mathrm{s}),[r]}}\right)^T \\ & + \sum_{m'=L_\mathrm{C}}^{m-\hat{L}_H} \mathbf{H}_{f_\mathrm{O},\hat{\mathbf{x}}_{m-m'}^{(\mathrm{s})}} \hat{\boldsymbol{\Sigma}}_{\breve{\mathbf{x}}_{m-m'}^{(\mathrm{s})}} \left(\mathbf{H}_{f_\mathrm{O},\hat{\mathbf{x}}_{m-m'}^{(\mathrm{s})}}\right)^T + \boldsymbol{\Sigma}_{\hat{\mathbf{v}}^{(\mathrm{s})}}\end{aligned} \tag{5.215}$$

wobei

$$\mathbf{H}_{f_\mathrm{O},\hat{\mathbf{z}}_{m|m,i}^{(\mathrm{s}),[r]}} = \begin{bmatrix}\mathbf{H}_{f_\mathrm{O},\hat{\boldsymbol{\chi}}_{m|m,i}^{(\mathrm{s}),[r]}} & \mathbf{H}_{f_\mathrm{O},\hat{\mathbf{n}}_{m|m,i}^{(\mathrm{s}),[r]}}\end{bmatrix} \tag{5.216}$$

mit

$$\mathbf{H}_{f_\mathrm{O},\hat{\boldsymbol{\chi}}_{m|m,i}^{(\mathrm{s}),[r]}} := \left.\frac{\partial f_\mathrm{O}\left(\boldsymbol{\chi}_m^{(\mathrm{s})}, \hat{\mathbf{x}}_{m-L_\mathrm{C}:m-\hat{L}_H}^{(\mathrm{s})}, \hat{\boldsymbol{\mu}}_{\breve{\mathbf{h}}_{0:\hat{L}_H}}, \hat{\mathbf{n}}_{m|m,i}^{(\mathrm{s}),[r]}\right)}{\partial \boldsymbol{\chi}_m^{(\mathrm{s})}}\right|_{\boldsymbol{\chi}_m^{(\mathrm{s})}=\hat{\boldsymbol{\chi}}_{m|m,i}^{(\mathrm{s}),[r]}}, \tag{5.217}$$

$$\mathbf{H}_{f_\mathrm{O},\hat{\mathbf{n}}_{m|m,i}^{(\mathrm{s}),[r]}} := \left.\frac{\partial f_\mathrm{O}\left(\hat{\boldsymbol{\chi}}_{m|m,i}^{(\mathrm{s}),[r]}, \hat{\mathbf{x}}_{m-L_\mathrm{C}:m-\hat{L}_H}^{(\mathrm{s})}, \hat{\boldsymbol{\mu}}_{\breve{\mathbf{h}}_{0:\hat{L}_H}}, \mathbf{n}_m^{(\mathrm{s})}\right)}{\partial \mathbf{n}_m^{(\mathrm{s})}}\right|_{\mathbf{n}_m^{(\mathrm{s})}=\hat{\mathbf{n}}_{m|m,i}^{(\mathrm{s}),[r]}}, \tag{5.218}$$

$$\mathbf{H}_{f_\mathrm{O},\hat{\mathbf{x}}_{m-m'}^{(\mathrm{s})}} := \left.\frac{\partial f_\mathrm{O}\left(\hat{\boldsymbol{\chi}}_{m|m,i}^{(\mathrm{s}),[r]}, \hat{\mathbf{x}}_{m-L_\mathrm{C}}^{(\mathrm{s})}, \ldots, \mathbf{x}_{m-m'}^{(\mathrm{s})}, \ldots, \hat{\mathbf{x}}_{m-\hat{L}_H}^{(\mathrm{s})}, \hat{\boldsymbol{\mu}}_{\breve{\mathbf{h}}_{0:\hat{L}_H}}, \hat{\mathbf{n}}_{m|m,i}^{(\mathrm{s}),[r]}\right)}{\partial \mathbf{x}_{m-m'}^{(\mathrm{s})}}\right|_{\mathbf{x}_{m-m'}^{(\mathrm{s})}=\hat{\mathbf{x}}_{m-m'}^{(\mathrm{s})}}. \tag{5.219}$$

Algorithmus 6 Berechnung der ersten beiden zentralen Momente der Verteilungsdichtefunktion $p\left(\mathbf{y}_m^{(\mathrm{s})} \middle| \hat{\mathbf{z}}_{m|m,i}^{(\mathrm{s}),[r]}, \mathbf{y}_{m-L_\mathrm{C}}^{(\mathrm{s})}, \hat{\mathbf{n}}_{m-L_\mathrm{C}}^{(\mathrm{s})}, \hat{\boldsymbol{\Sigma}}_{\check{\mathbf{n}}_{m-L_\mathrm{C}}^{(\mathrm{s})}}, \hat{\boldsymbol{\mu}}_{\check{\mathbf{h}}_{0:L_\mathrm{C}-1}}\right)$ basierend auf der rekursiven Beobachtungsfunktion (5.185).

Eingabe: $\hat{\mathbf{z}}_{m|m,i}^{(\mathrm{s}),[r]}$, $\mathbf{y}_{m-L_\mathrm{C}}^{(\mathrm{s})}$, $\hat{\mathbf{n}}_{m-L_\mathrm{C}}^{(\mathrm{s})}$, $\hat{\boldsymbol{\Sigma}}_{\check{\mathbf{n}}_{m-L_\mathrm{C}}^{(\mathrm{s})}}$, $\hat{\boldsymbol{\mu}}_{\check{\mathbf{h}}_{0:L_\mathrm{C}-1}}$.

Ausgabe: $\hat{\mathbf{y}}_{m,i}^{(\mathrm{s}),[r]}$, $\hat{\boldsymbol{\Sigma}}_{\mathbf{y}_{m,i}^{(\mathrm{s}),[r]}}$, $\mathbf{H}_{f_{\mathrm{O},L_\mathrm{C}}^{(\mathrm{R})},\hat{\mathbf{z}}_{m|m,i}^{(\mathrm{s}),[r]}}$.

- Berechne die prädizierte Beobachtung $\hat{\mathbf{y}}_{m,i}^{(\mathrm{s}),[r]}$ und die zugehörige Kovarianzmatrix $\hat{\boldsymbol{\Sigma}}_{\mathbf{y}_{m,i}^{(\mathrm{s}),[r]}}$ basierend auf der aktuellen Linearisierungsstelle $\hat{\mathbf{z}}_{m|m,i}^{(\mathrm{s}),[r]}$ gemäß

$$\hat{\mathbf{y}}_{m,i}^{(\mathrm{s}),[r]} = f_{\mathrm{O},L_\mathrm{C}}^{(\mathrm{R})}\left(\hat{\boldsymbol{\chi}}_{m|m,i}^{(\mathrm{s}),[r]}, \hat{\boldsymbol{\mu}}_{\check{\mathbf{h}}_{0:L_\mathrm{C}-1}}, \mathbf{y}_{m-L_\mathrm{C}}^{(\mathrm{s})}, \hat{\mathbf{n}}_{m|m,i}^{(\mathrm{s}),[r]}, \hat{\mathbf{n}}_{m-L_\mathrm{C}}^{(\mathrm{s})}\right) + \boldsymbol{\mu}_{\hat{\mathbf{v}}_{L_\mathrm{C}}^{(\mathrm{s,R})}} \tag{5.220}$$

$$\begin{aligned}\hat{\boldsymbol{\Sigma}}_{\mathbf{y}_{m,i}^{(\mathrm{s}),[r]}} = \mathbf{H}_{f_{\mathrm{O},L_\mathrm{C}}^{(\mathrm{R})},\hat{\boldsymbol{\chi}}_{m|m,i}^{(\mathrm{s}),[r]}} \hat{\boldsymbol{\Sigma}}_{\mathbf{z}_{m|m-1,i}^{(\mathrm{s})}} \left(\mathbf{H}_{f_{\mathrm{O},L_\mathrm{C}}^{(\mathrm{R})},\hat{\boldsymbol{\chi}}_{m|m,i}^{(\mathrm{s}),[r]}}\right)^T \\ + \mathbf{H}_{f_{\mathrm{O},L_\mathrm{C}}^{(\mathrm{R})},\hat{\mathbf{n}}_{m-L_\mathrm{C}}^{(\mathrm{s})}} \hat{\boldsymbol{\Sigma}}_{\check{\mathbf{n}}_{m-L_\mathrm{C}}^{(\mathrm{s})}} \left(\mathbf{H}_{f_{\mathrm{O},L_\mathrm{C}}^{(\mathrm{R})},\hat{\mathbf{n}}_{m-L_\mathrm{C}}^{(\mathrm{s})}}\right)^T + \boldsymbol{\Sigma}_{\hat{\mathbf{v}}_{L_\mathrm{C}}^{(\mathrm{s,R})}},\end{aligned} \tag{5.221}$$

wobei

$$\mathbf{H}_{f_{\mathrm{O},L_\mathrm{C}}^{(\mathrm{R})},\hat{\mathbf{z}}_{m|m,i}^{(\mathrm{s}),[r]}} = \left[\mathbf{H}_{f_{\mathrm{O},L_\mathrm{C}}^{(\mathrm{R})},\hat{\boldsymbol{\chi}}_{m|m,i}^{(\mathrm{s}),[r]}} \quad \mathbf{H}_{f_{\mathrm{O},L_\mathrm{C}}^{(\mathrm{R})},\hat{\mathbf{n}}_{m|m,i}^{(\mathrm{s}),[r]}}\right] \tag{5.222}$$

mit

$$\mathbf{H}_{f_{\mathrm{O},L_\mathrm{C}}^{(\mathrm{R})},\hat{\boldsymbol{\chi}}_{m|m,i}^{(\mathrm{s}),[r]}} := \left.\frac{\partial f_{\mathrm{O},L_\mathrm{C}}^{(\mathrm{R})}\left(\boldsymbol{\chi}_m^{(\mathrm{s})}, \hat{\boldsymbol{\mu}}_{\check{\mathbf{h}}_{0:L_\mathrm{C}-1}}, \mathbf{y}_{m-L_\mathrm{C}}^{(\mathrm{s})}, \hat{\mathbf{n}}_{m|m,i}^{(\mathrm{s}),[r]}, \hat{\mathbf{n}}_{m-L_\mathrm{C}}^{(\mathrm{s})}\right)}{\partial \boldsymbol{\chi}_m^{(\mathrm{s})}}\right|_{\boldsymbol{\chi}_m^{(\mathrm{s})}=\hat{\boldsymbol{\chi}}_{m|m,i}^{(\mathrm{s}),[r]}}, \tag{5.223}$$

$$\mathbf{H}_{f_{\mathrm{O},L_\mathrm{C}}^{(\mathrm{R})},\hat{\mathbf{n}}_{m|m,i}^{(\mathrm{s}),[r]}} := \left.\frac{\partial f_{\mathrm{O},L_\mathrm{C}}^{(\mathrm{R})}\left(\hat{\boldsymbol{\chi}}_{m|m,i}^{(\mathrm{s}),[r]}, \hat{\boldsymbol{\mu}}_{\check{\mathbf{h}}_{0:L_\mathrm{C}-1}}, \mathbf{y}_{m-L_\mathrm{C}}^{(\mathrm{s})}, \mathbf{n}_m^{(\mathrm{s})}, \hat{\mathbf{n}}_{m-L_\mathrm{C}}^{(\mathrm{s})}\right)}{\partial \mathbf{n}_m^{(\mathrm{s})}}\right|_{\mathbf{n}_m^{(\mathrm{s})}=\hat{\mathbf{n}}_{m|m,i}^{(\mathrm{s}),[r]}}, \tag{5.224}$$

$$\mathbf{H}_{f_{\mathrm{O},L_\mathrm{C}}^{(\mathrm{R})},\hat{\mathbf{n}}_{m-L_\mathrm{C}+1}^{(\mathrm{s})}} := \left.\frac{\partial f_{\mathrm{O},L_\mathrm{C}}^{(\mathrm{R})}\left(\hat{\boldsymbol{\chi}}_{m|m,i}^{(\mathrm{s}),[r]}, \hat{\boldsymbol{\mu}}_{\check{\mathbf{h}}_{0:L_\mathrm{C}-1}}, \mathbf{y}_{m-L_\mathrm{C}}^{(\mathrm{s})}, \hat{\mathbf{n}}_{m|m,i}^{(\mathrm{s}),[r]}, \mathbf{n}_{m-L_\mathrm{C}}^{(\mathrm{s})}\right)}{\partial \mathbf{n}_{m-L_\mathrm{C}}^{(\mathrm{s})}}\right|_{\mathbf{n}_{m-L_\mathrm{C}}^{(\mathrm{s})}=\hat{\mathbf{n}}_{m-L_\mathrm{C}}^{(\mathrm{s})}}. \tag{5.225}$$

5.3.2. Modellkombinationsalgorithmen

Wird, wie im vorhergehenden Abschnitt, unter der Annahme einer GAUSS-förmigen A-posteriori-Verteilungsdichtefunktion $p\left(\mathbf{z}_{m-1}^{(\mathrm{s})}\middle|\mathbf{y}_{1:m-1}^{(\mathrm{s})}\right)$ zum Zeitpunkt $m-1$ die auf das i-te Teilmodell bedingte A-posteriori-Verteilungsdichtefunktion $p\left(\mathbf{z}_{m}^{(\mathrm{s})}\middle|\mathbf{y}_{1:m}^{(\mathrm{s})},\zeta_m=i\right)$ zum Zeitpunkt m durch eine GAUSS-Verteilungsdichtefunktion approximiert, so folgt für die A-posteriori-Verteilungsdichtefunktion $p\left(\mathbf{z}_{m}^{(\mathrm{s})}\middle|\mathbf{y}_{1:m}^{(\mathrm{s})}\right)$ zwangsläufig, dass sie durch ein *GMM* gemäß

$$p\left(\mathbf{z}_{m}^{(\mathrm{s})}\middle|\mathbf{y}_{1:m}^{(\mathrm{s})}\right)\approx\sum_{i=1}^{I}P\left(\zeta_m=i\middle|\mathbf{y}_{1:m}^{(\mathrm{s})}\right)\mathcal{N}\left(\mathbf{z}_{m}^{(\mathrm{s})};\hat{\mathbf{z}}_{m|m,i}^{(\mathrm{s})},\hat{\boldsymbol{\Sigma}}_{\mathbf{z}_{m|m,i}^{(\mathrm{s})}}\right)\tag{5.226}$$

repräsentiert wird. Um die A-posteriori-Verteilungsdichtefunktion zum Zeitpunkt $m+1$ zu bestimmen, ließe sich das zuvor beschriebene Prinzip auf jede Mischungskomponente getrennt anwenden, so dass die Approximation

$$\begin{aligned}p\left(\mathbf{z}_{m+1}^{(\mathrm{s})}\middle|\mathbf{y}_{1:m+1}^{(\mathrm{s})}\right)\approx\sum_{i=1}^{I}\sum_{k=1}^{I}P\left(\zeta_m=i,\zeta_{m+1}=k\middle|\mathbf{y}_{1:m+1}^{(\mathrm{s})}\right)\\\cdot\mathcal{N}\left(\mathbf{z}_{m+1}^{(\mathrm{s})};\hat{\mathbf{z}}_{m+1|m+1,i,k}^{(\mathrm{s})},\hat{\boldsymbol{\Sigma}}_{\mathbf{z}_{m+1|m+1,i,k}^{(\mathrm{s})}}\right)\end{aligned}\tag{5.227}$$

resultiert. Anhand dieses Beispiels lässt sich erkennen, dass die Anzahl der Mischungskomponenten zur Darstellung der A-posteriori-Verteilungsdichtefunktion, und damit auch der Rechenaufwand, exponentiell mit dem Segmentindex wächst. Um diesem Phänomen entgegenzuwirken, werden hier drei mögliche Verfahren aus der Literatur vorgestellt. Darunter befinden sich die sogenannte generalisierte pseudo-BAYES'sche Schätzung (engl. ***G**eneralized **P**seudo* **B**AYESIAN (*GPB*) estimation) erster und zweiter Ordnung sowie die Schätzung mit interagierenden Modellen (engl. ***I**nteracting **M**ultiple **M**odel* (*IMM*) estimation) [BSLK01].

Bei der *GPB*-Schätzung erster Ordnung (engl. ***G**eneralized **P**seudo* **B**AYESIAN *estimation of order **1*** (*GPB1*)), die ausführlich in Alg. 7 beschrieben ist, wird die A-posteriori-Verteilungsdichtefunktion nach jedem Inferenzschritt durch eine GAUSS-Verteilungsdichtefunktion approximiert. Der Mittelwertvektor $\hat{\mathbf{z}}_{m|m}^{(\mathrm{s})}$ und die Kovarianzmatrix $\hat{\boldsymbol{\Sigma}}_{\mathbf{z}_{m|m}^{(\mathrm{s})}}$ der A-posteriori-Verteilungsdichtefunktion

$$p\left(\mathbf{z}_{m}^{(\mathrm{s})}\middle|\mathbf{y}_{1:m}^{(\mathrm{s})}\right)\approx\mathcal{N}\left(\mathbf{z}_{m}^{(\mathrm{s})};\hat{\mathbf{z}}_{m|m}^{(\mathrm{s})},\hat{\boldsymbol{\Sigma}}_{\mathbf{z}_{m|m}^{(\mathrm{s})}}\right)\tag{5.228}$$

werden dabei derart bestimmt, dass die KULLBACK-LEIBLER-Divergenz zwischen (5.228) und dem *GMM* (5.226) minimiert wird. Daraus ergibt sich die Modellkombinationsvorschrift gemäß (5.233) und (5.234).

Die *IMM*-Schätzung, aufgeführt in Alg. 8, basiert auf der Darstellung der auf das i-te Teilmodell bedingten A-priori-Verteilungsdichtefunktion zum Zeitpunkt m gemäß

$$p\left(\mathbf{z}_{m}^{(\mathrm{s})}\middle|\mathbf{y}_{1:m-1}^{(\mathrm{s})},\zeta_m=i\right)=\sum_{k=1}^{I}P\left(\zeta_{m-1}=k\middle|\zeta_m=i,\mathbf{y}_{1:m-1}^{(\mathrm{s})}\right)p\left(\mathbf{z}_{m}^{(\mathrm{s})}\middle|\mathbf{y}_{1:m-1}^{(\mathrm{s})},\zeta_m=i,\zeta_{m-1}=k\right)\tag{5.237}$$

Algorithmus 7 Modellkombination gemäß *GPB1*

- **Initialisierung:**
  Initialisiere die Schätzwerte für die Mittelwertvektoren und Kovarianzmatrizen der Merkmalsvektoren der sauberen Sprachsignals und des Störsignals für $m \in \{-\hat{L}_H+1,\ldots,0\}$ durch

$$\hat{\mathbf{x}}_m^{(\mathrm{s})} = (x_{\mathrm{MIN}},\ldots,x_{\mathrm{MIN}})^T, \quad \hat{\boldsymbol{\Sigma}}_{\breve{\mathbf{x}}_m^{(\mathrm{s})}} = \sigma_{\mathrm{MIN}}^2 \cdot \mathbf{I}, \quad \hat{\mathbf{n}}_m^{(\mathrm{s})} = \boldsymbol{\mu}_{\mathbf{n}}, \quad \hat{\boldsymbol{\Sigma}}_{\breve{\mathbf{n}}_m^{(\mathrm{s})}} = \sigma_{\mathrm{MIN}}^2 \cdot \mathbf{I} \tag{5.229}$$

  sowie für $m \in \{-L_{\mathrm{C}}+1,\ldots,0\}$ die Merkmalsvektoren $\mathbf{y}_m^{(\mathrm{s})}$ durch $\mathbf{y}_m^{(\mathrm{s})} = \hat{\mathbf{x}}_m^{(\mathrm{s})}$.
- **Filterung:**
  Für $m = 1..M$
  - **Modellabhängige Inferenzen**:
    1. Berechne die A-priori-Modell-WSKs $P_{m|m-1,i} := P\left(\zeta_m = i|\mathbf{y}_{1:m-1}^{(\mathrm{s})}\right)$ für $i \in \{1,\ldots,I\}$:
       Wenn $m = 1$ **dann**

$$P_{m|m-1,i} = \psi_i, \tag{5.230}$$

       Sonst

$$P_{m|m-1,i} = \sum_{k=1}^{I} a_{k,i} P_{m-1|m-1,k}. \tag{5.231}$$

       Ende wenn
    2. Wende für $i \in \{1,\ldots,I\}$ das *IEKF* gemäß Alg. 4 an:
       Eingabe: $\hat{\mathbf{z}}_{m-1|m-1}^{(\mathrm{s})}$, $\hat{\boldsymbol{\Sigma}}_{\mathbf{z}_{m-1|m-1}^{(\mathrm{s})}}$, $\hat{\mathbf{x}}_{m-\hat{L}_H:m-L_{\mathrm{C}}}^{(\mathrm{s})}$, $\hat{\boldsymbol{\Sigma}}_{\breve{\mathbf{x}}_{m-\hat{L}_H:m-L_{\mathrm{C}}}^{(\mathrm{s})}}$, $\hat{\mathbf{n}}_{m-L_{\mathrm{C}}}^{(\mathrm{s})}$, $\hat{\boldsymbol{\Sigma}}_{\breve{\mathbf{n}}_{m-L_{\mathrm{C}}}^{(\mathrm{s})}}$, $\hat{\mathbf{x}}_{-L_{\mathrm{C}}+2:0}^{(\mathrm{s})}$, $\hat{\boldsymbol{\Sigma}}_{\breve{\mathbf{x}}_{-L_{\mathrm{C}}+2:0}^{(\mathrm{s})}}$, $\mathbf{y}_{m-L_{\mathrm{C}}}^{(\mathrm{s})}$, $\hat{\boldsymbol{\mu}}_{\breve{\mathbf{h}}_{0:\hat{L}_H}}$, m, k.
       Ausgabe: $\hat{\mathbf{z}}_{m|m,i}^{(\mathrm{s})}$, $\hat{\boldsymbol{\Sigma}}_{\mathbf{z}_{m|m,i}^{(\mathrm{s})}}$, $\hat{\mathbf{y}}_{m,i}^{(\mathrm{s}),[1]}$, $\hat{\boldsymbol{\Sigma}}_{\mathbf{y}_{m,i}^{(\mathrm{s}),[1]}}$.
  - **Modellkombination**:
    1. Berechne für $i \in \{1,\ldots,I\}$ die A-posteriori-Modell-WSKs $P_{m|m,i} := P\left(\zeta_m = i|\mathbf{y}_{1:m}^{(\mathrm{s})}\right)$:

$$P_{m|m,i} \propto \mathcal{N}\left(\mathbf{y}_m^{(\mathrm{s})}; \hat{\mathbf{y}}_{m,i}^{(\mathrm{s}),[1]}; \hat{\boldsymbol{\Sigma}}_{\mathbf{y}_{m,i}^{(\mathrm{s}),[1]}}\right) P_{m|m-1,i}. \tag{5.232}$$

    2. Berechne den Mittelwertvektor und die Kovarianzmatrix der A-posteriori-Verteilungsdichtefunktion $p\left(\mathbf{z}_m^{(\mathrm{s})} \middle| \mathbf{y}_{1:m}^{(\mathrm{s})}\right)$ gemäß

$$\hat{\mathbf{z}}_{m|m}^{(\mathrm{s})} = \sum_{i=1}^{I} P_{m|m,i} \hat{\mathbf{z}}_{m|m,i}^{(\mathrm{s})}, \tag{5.233}$$

$$\hat{\boldsymbol{\Sigma}}_{\mathbf{z}_{m|m}^{(\mathrm{s})}} = \sum_{i=1}^{I} P_{m|m,i} \left[\hat{\boldsymbol{\Sigma}}_{\mathbf{z}_{m|m,i}^{(\mathrm{s})}} + \left(\hat{\mathbf{z}}_{m|m,i}^{(\mathrm{s})} - \hat{\mathbf{z}}_{m|m}^{(\mathrm{s})}\right)\left(\hat{\mathbf{z}}_{m|m,i}^{(\mathrm{s})} - \hat{\mathbf{z}}_{m|m}^{(\mathrm{s})}\right)^T\right]. \tag{5.234}$$

  - **Extraktion der Schätzungen**:
    1. Extrahiere den geschätzten Merkmalsvektor der Störung sowie zugehörige Schätzfehlerkovarianzmatrix aus dem Zustandsvektor und der Zustandskovarianzmatrix:

$$\hat{\mathbf{n}}_m^{(\mathrm{s})} = \mathbf{M}_{\mathbf{n},\mathrm{EXTR}} \hat{\mathbf{z}}_{m|m}^{(\mathrm{s})}, \quad \hat{\boldsymbol{\Sigma}}_{\breve{\mathbf{n}}_m^{(\mathrm{s})}} = \mathbf{M}_{\mathbf{n},\mathrm{EXTR}} \hat{\boldsymbol{\Sigma}}_{\mathbf{z}_{m|m}^{(\mathrm{s})}} \left(\mathbf{M}_{\mathbf{n},\mathrm{EXTR}}\right)^T \tag{5.235}$$

       mit $\mathbf{M}_{\mathbf{n},\mathrm{EXTR}} := \begin{bmatrix}\mathbf{0} & \ldots & \mathbf{0} & \mathbf{I}\end{bmatrix} \in \mathbb{R}^{Q\times(L_{\mathrm{C}}+1)Q}$.
    2. **Wenn** $m \geq L_{\mathrm{C}}$ **dann**
       - Extrahiere den verbesserten Merkmalsvektor samt der Schätzfehlerkovarianzmatrix aus dem Zustandsvektor und der Zustandskovarianzmatrix:

$$\hat{\mathbf{x}}_{m-L_{\mathrm{C}}+1}^{(\mathrm{s})} = \mathbf{M}_{\mathbf{x},\mathrm{EXTR}} \hat{\mathbf{z}}_{m|m}^{(\mathrm{s})}, \quad \hat{\boldsymbol{\Sigma}}_{\breve{\mathbf{x}}_{m-L_{\mathrm{C}}+1}^{(\mathrm{s})}} = \mathbf{M}_{\mathbf{x},\mathrm{EXTR}} \hat{\boldsymbol{\Sigma}}_{\mathbf{z}_{m|m}^{(\mathrm{s})}} \left(\mathbf{M}_{\mathbf{x},\mathrm{EXTR}}\right)^T \tag{5.236}$$

         mit $\mathbf{M}_{\mathbf{x},\mathrm{EXTR}} := \begin{bmatrix}\mathbf{0} & \ldots & \mathbf{0} & \mathbf{I} & \mathbf{0}\end{bmatrix} \in \mathbb{R}^{Q\times(L_{\mathrm{C}}+1)Q}$.

       Ende wenn

  Ende für

und einer Approximation von $p\left(\mathbf{z}_m^{(\mathrm{s})}\middle|\mathbf{y}_{1:m-1}^{(\mathrm{s})},\zeta_m=i,\zeta_{m-1}=k\right)$ durch eine GAUSS-Verteilungsdichtefunktion. Daher wird hierbei im Gegensatz zur *GPB1*-Schätzung zum Zeitpunkt m für jedes Teilmodell i das *IEKF* auf Grundlage eines teilmodellspezifischen initialen Mittelwertvektors $\hat{\mathbf{z}}_{m-1,i}^{(\mathrm{s,INIT})}$ und einer Kovarianzmatrix $\hat{\boldsymbol{\Sigma}}_{m-1,i}^{(\mathrm{s,INIT})}$ ausgeführt.

Bei der *GPB*-Schätzung zweiter Ordnung (engl. ***G**eneralized **P**seudo* **B**AYESIAN *estimation of order **2*** (*GPB2*)), welche in Alg. 9 dargestellt ist, findet nach jedem Inferenzschritt eine Approximation der A-posteriori-Verteilungsdichtefunktion durch ein *GMM* mit I Mischungskomponenten gemäß (5.226) statt. Die Anzahl der erforderlichen Aufrufe des *IEKF* pro Inferenzschritt ist daher I^2 im Vergleich zu I bei der *GPB1*- und *IMM*-Schätzung.

Die Initialisierung ist bei allen drei Verfahren identisch. Unter der Annahme, dass für einen Dauer von $\hat{L}_H-1$ Segmenten unmittelbar vor dem Zeitpunkt $m=1$ keine Sprache im Signal auftritt und das Störsignal stationär ist, lässt sich diese gemäß (5.229) bewerkstelligen. Vernünftige Werte für die Parameter x_{MIN} und σ_{MIN}^2 sind beispielsweise $x_{\mathrm{MIN}}=-50$ und $\sigma_{\mathrm{MIN}}^2=10^{-6}$.

Weiterhin muss bemerkt werden, dass im Sinne der Gewinnungen von Punktschätzungen $\hat{\mathbf{z}}_{m|m}^{(\mathrm{s})}$ und zugehörigen Schätzfehlerkovarianzmatrizen $\hat{\boldsymbol{\Sigma}}_{\mathbf{z}_{m|m}^{(\mathrm{s})}}$ bei allen drei Verfahren die A-posteriori-Verteilungsdichtefunktion gemäß (5.228) angenähert werden muss. Aus diesen Schätzungen werden abschließend mit Hilfe von (5.235) und (5.236) die verbesserten Merkmale $\hat{\mathbf{x}}_{m-L_{\mathrm{C}}+1}^{(\mathrm{s})}$ und $\hat{\mathbf{n}}_m^{(\mathrm{s})}$ sowie die entsprechenden Schätzfehlerkovarianzmatrizen $\hat{\boldsymbol{\Sigma}}_{\breve{\mathbf{n}}_m^{(\mathrm{s})}}$ und $\hat{\boldsymbol{\Sigma}}_{\breve{\mathbf{x}}_{m-L_{\mathrm{C}}+1}^{(\mathrm{s})}}$ extrahiert.

Da die Schätzung $\hat{\mathbf{x}}_{m-L_{\mathrm{C}}+1}^{(\mathrm{s})}$ bedingt auf die Beobachtungen $\mathbf{y}_{1:m}^{(\mathrm{s})}$ ist, beinhaltet sie Information über einen gewissen Zeitraum der Dauer von $L_{\mathrm{C}}-1$ Segmenten in der Zukunft. Obwohl diese Art der impliziten Glättung eine Latenz der gleichen Dauer verursacht, sind die hier beschriebenen Verfahren in der Regel für eine Online-Verarbeitung geeignet, da die Werte von L_{C} relativ klein gewählt werden können.

Für weitere grundlegende Details zu den hier aufgeführten Modellkombinationsalgorithmen sei auf eine ausführliche Beschreibung in [BSLK01] verwiesen.

Algorithmus 8 Modellkombination gemäß *IMM*

- **Initialisierung:**
  1. Initialisiere $\hat{\mathbf{x}}_m^{(\mathrm{s})}$, $\hat{\mathbf{\Sigma}}_{\check{\mathbf{x}}_m^{(\mathrm{s})}}$, $\hat{\mathbf{n}}_m^{(\mathrm{s})}$ und $\hat{\mathbf{\Sigma}}_{\check{\mathbf{n}}_m^{(\mathrm{s})}}$ für $m \in \left\{-\hat{L}_H+1,\ldots,0\right\}$ gemäß (5.229).
  2. Initialisiere $\mathbf{y}_m^{(\mathrm{s})}$ für $m \in \{-L_\mathrm{C}+1,\ldots,0\}$ wie in Alg. 7.
  3. Initialisiere für $i \in \{1,\ldots,I\}$ den Zustandsvektor $\hat{\mathbf{z}}_{0|0,i}^{(\mathrm{s})}$ und die Kovarianzmatrix $\hat{\mathbf{\Sigma}}_{\mathbf{z}_{0|0,i}^{(\mathrm{s})}}$:

$$\hat{\mathbf{z}}_{0|0,i}^{(\mathrm{s})} = \left[\left(\hat{\mathbf{x}}_0^{(\mathrm{s})}\right)^T,\ldots,\left(\hat{\mathbf{x}}_{-L_\mathrm{C}+1}^{(\mathrm{s})}\right)^T,\left(\hat{\mathbf{n}}_0^{(\mathrm{s})}\right)^T\right]^T,\quad \hat{\mathbf{\Sigma}}_{\mathbf{z}_{0|0,i}^{(\mathrm{s})}} = \mathrm{blockdiag}\left\{\sigma_{\mathrm{MIN}}^2\mathbf{I},\ldots,\sigma_{\mathrm{MIN}}^2\mathbf{I}\right\}. \tag{5.238}$$

- **Filterung:**

  Für $m = 1..M$
  - **Modellabhängige Inferenzen**:
    1. Berechne für $i \in \{1,\ldots,I\}$ die A-priori-Modell-WSKs $P_{m|m-1,i}$ gemäß (5.230) und (5.231).
    2. Berechne für alle Tupel (i,k) mit $i,k \in \{1,\ldots,I\}$ die Mischungswahrscheinlichkeiten $P_{i,k,m}^{(\mathrm{MIX})} := P\left(\zeta_{m-1}=k\middle|\zeta_m=i,\mathbf{y}_{1:m-1}^{(\mathrm{s})}\right)$ gemäß

$$P_{i,k,m}^{(\mathrm{MIX})} \propto a_{k,i}P_{m-1|m-1,i}. \tag{5.239}$$

    3. Berechne für $i \in \{1,\ldots,I\}$ die initialen Mittelwertvektoren und Kovarianzmatrizen für das i-te *IEKF* gemäß

$$\hat{\mathbf{z}}_{m-1,i}^{(\mathrm{s,INIT})} = \sum_{k=1}^{I} P_{i,k,m}^{(\mathrm{MIX})}\hat{\mathbf{z}}_{m-1|m-1,k}^{(\mathrm{s})}, \tag{5.240}$$

$$\hat{\mathbf{\Sigma}}_{m-1,i}^{(\mathrm{s,INIT})} = \sum_{k=1}^{I} P_{i,k,m}^{(\mathrm{MIX})}\left[\hat{\mathbf{\Sigma}}_{\mathbf{z}_{m-1|m-1,k}^{(\mathrm{s})}} + \left(\hat{\mathbf{z}}_{m-1|m-1,k}^{(\mathrm{s})} - \hat{\mathbf{z}}_{m-1,i}^{(\mathrm{s,INIT})}\right)\left(\hat{\mathbf{z}}_{m-1|m-1,k}^{(\mathrm{s})} - \hat{\mathbf{z}}_{m-1,i}^{(\mathrm{s,INIT})}\right)^T\right]. \tag{5.241}$$

    4. Wende für $i \in \{1,\ldots,I\}$ das *IEKF* gemäß Alg. 4 an:

       Eingabe: $\hat{\mathbf{z}}_{m-1,i}^{(\mathrm{s,INIT})}$, $\hat{\mathbf{\Sigma}}_{m-1,i}^{(\mathrm{s,INIT})}$, $\hat{\mathbf{x}}_{m-\hat{L}_H:m-L_\mathrm{C}}^{(\mathrm{s})}$, $\hat{\mathbf{\Sigma}}_{\check{\mathbf{x}}_{m-\hat{L}_H:m-L_\mathrm{C}}^{(\mathrm{s})}}$, $\hat{\mathbf{n}}_{m-L_\mathrm{C}}^{(\mathrm{s})}$, $\hat{\mathbf{\Sigma}}_{\check{\mathbf{n}}_{m-L_\mathrm{C}}^{(\mathrm{s})}}$, $\hat{\mathbf{x}}_{-L_\mathrm{C}+2:0}^{(\mathrm{s})}$, $\hat{\mathbf{\Sigma}}_{\check{\mathbf{x}}_{-L_\mathrm{C}+2:0}^{(\mathrm{s})}}$, $\mathbf{y}_{m-L_\mathrm{C}}^{(\mathrm{s})}$, $\hat{\boldsymbol{\mu}}_{\check{\mathbf{h}}_{0:\hat{L}_H}}$, m, i.

       Ausgabe: $\hat{\mathbf{z}}_{m|m,i}^{(\mathrm{s})}$, $\hat{\mathbf{\Sigma}}_{\mathbf{z}_{m|m,i}^{(\mathrm{s})}}$, $\hat{\mathbf{y}}_{m,i}^{(\mathrm{s}),[1]}$, $\hat{\mathbf{\Sigma}}_{\mathbf{y}_{m,i}^{(\mathrm{s}),[1]}}$.
  - **Modellkombination**:
    1. Berechne für $i \in \{1,\ldots,I\}$ die A-posteriori-Modell-WSKs $P_{m|m,i}$ gemäß (5.232).
    2. Berechne den Mittelwertvektor und die Kovarianzmatrix der A-posteriori-Verteilungsdichtefunktion $p\left(\mathbf{z}_m^{(\mathrm{s})}\middle|\mathbf{y}_{1:m}^{(\mathrm{s})}\right)$ gemäß

$$\hat{\mathbf{z}}_{m|m}^{(\mathrm{s})} = \sum_{i=1}^{I} P_{m|m,i}\hat{\mathbf{z}}_{m|m,i}^{(\mathrm{s})}, \tag{5.242}$$

$$\hat{\mathbf{\Sigma}}_{\mathbf{z}_{m|m}^{(\mathrm{s})}} = \sum_{i=1}^{I} P_{m|m,i}\left[\hat{\mathbf{\Sigma}}_{\mathbf{z}_{m|m,i}^{(\mathrm{s})}} + \left(\hat{\mathbf{z}}_{m|m,i}^{(\mathrm{s})} - \hat{\mathbf{z}}_{m|m}^{(\mathrm{s})}\right)\left(\hat{\mathbf{z}}_{m|m,i}^{(\mathrm{s})} - \hat{\mathbf{z}}_{m|m}^{(\mathrm{s})}\right)^T\right]. \tag{5.243}$$

  - **Extraktion der Schätzungen**:
    1. Extrahiere $\hat{\mathbf{n}}_m^{(\mathrm{s})}$ und $\hat{\mathbf{\Sigma}}_{\check{\mathbf{n}}_m^{(\mathrm{s})}}$ aus $\hat{\mathbf{z}}_{m|m}^{(\mathrm{s})}$ und $\hat{\mathbf{\Sigma}}_{\mathbf{z}_{m|m}^{(\mathrm{s})}}$ gemäß (5.235).
    2. **Wenn** $m \geq L_\mathrm{C}$ **dann**
       - Extrahiere $\hat{\mathbf{x}}_{m-L_\mathrm{C}+1}^{(\mathrm{s})}$ und $\hat{\mathbf{\Sigma}}_{\check{\mathbf{x}}_{m-L_\mathrm{C}+1}^{(\mathrm{s})}}$ aus $\hat{\mathbf{z}}_{m|m}^{(\mathrm{s})}$ und $\hat{\mathbf{\Sigma}}_{\mathbf{z}_{m|m}^{(\mathrm{s})}}$ gemäß (5.236).

       Ende wenn

Ende für

Algorithmus 9 Modellkombination gemäß *GPB2*

- **Initialisierung:**
  1. Initialisiere $\hat{\mathbf{x}}_m^{(\mathrm{s})}$, $\hat{\mathbf{\Sigma}}_{\check{\mathbf{x}}_m^{(\mathrm{s})}}$, $\hat{\mathbf{n}}_m^{(\mathrm{s})}$ und $\hat{\mathbf{\Sigma}}_{\check{\mathbf{n}}_m^{(\mathrm{s})}}$ für $m \in \left\{-\hat{L}_H+1,\ldots,0\right\}$ gemäß (5.229).
  2. Initialisiere $\mathbf{y}_m^{(\mathrm{s})}$ für $m \in \{-L_\mathrm{C}+1,\ldots,0\}$ wie in Alg. 7.
  3. Initialisiere $\hat{\mathbf{z}}_{0|0,i}^{(\mathrm{s})}$ und $\hat{\mathbf{\Sigma}}_{\mathbf{z}_{0|0,i}^{(\mathrm{s})}}$ für $i \in \{1,\ldots,I\}$ gemäß (5.238).
  4. Initialisiere für $i \in \{1,\ldots,I\}$ die A-posteriori-Modell-WSKs durch $P_{0|0,i} = \psi_i$.
- **Filterung:**

  Für $m = 1..M$
  - **Modellabhängige Inferenzen**:
    1. Wende für alle Tupel (i,k) mit $i,k \in \{1,\ldots,I\}$ das *IEKF* gemäß Alg. 4 an:

       Eingabe: $\hat{\mathbf{z}}_{m-1|m-1,i}^{(\mathrm{s})}$, $\hat{\mathbf{\Sigma}}_{\mathbf{z}_{m-1|m-1,i}^{(\mathrm{s})}}$, $\hat{\mathbf{x}}_{m-\hat{L}_H:m-L_\mathrm{C}}^{(\mathrm{s})}$, $\hat{\mathbf{\Sigma}}_{\check{\mathbf{x}}_{m-\hat{L}_H:m-L_\mathrm{C}}^{(\mathrm{s})}}$, $\hat{\mathbf{n}}_{m-L_\mathrm{C}}^{(\mathrm{s})}$, $\hat{\mathbf{\Sigma}}_{\check{\mathbf{n}}_{m-L_\mathrm{C}}^{(\mathrm{s})}}$, $\hat{\mathbf{x}}_{-L_\mathrm{C}+2:0}^{(\mathrm{s})}$, $\hat{\mathbf{\Sigma}}_{\check{\mathbf{x}}_{-L_\mathrm{C}+2:0}^{(\mathrm{s})}}$, $\mathbf{y}_{m-L_\mathrm{C}}^{(\mathrm{s})}$, $\hat{\boldsymbol{\mu}}_{\check{\mathbf{h}}_{0:\hat{L}_H}}$, m, k.

       Ausgabe: $\hat{\mathbf{z}}_{m|m,i,k}^{(\mathrm{s})}$, $\hat{\mathbf{\Sigma}}_{\mathbf{z}_{m|m,i,k}^{(\mathrm{s})}}$, $\hat{\mathbf{y}}_{m,i,k}^{(\mathrm{s}),[1]}$, $\hat{\mathbf{\Sigma}}_{\mathbf{y}_{m,i,k}^{(\mathrm{s}),[1]}}$.
  - **Modellkombination**:
    1. Berechne für alle Tupel (i,k) mit $i,k \in \{1,\ldots,I\}$ die Fusions-WSKs $P_{m,k,i}^{(\mathrm{FUS})} := P\left(\zeta_{m-1}=k \middle| \zeta_m = i, \mathbf{y}_{1:m}^{(\mathrm{s})}\right)$ gemäß

       $$P_{m,k,i}^{(\mathrm{FUS})} \propto \mathcal{N}\left(\mathbf{y}_m^{(\mathrm{s})}; \hat{\mathbf{y}}_{m,i,k}^{(\mathrm{s}),[1]}, \hat{\mathbf{\Sigma}}_{\mathbf{y}_{m,i,k}^{(\mathrm{s}),[1]}}\right) a_{k,i} P_{m-1|m-1,k} \tag{5.244}$$

    2. Berechne für $k \in \{1,\ldots,I\}$ den Mittelwertvektor und die Kovarianzmatrix der modellbedingten A-posteriori-Verteilungsdichtefunktion $p\left(\mathbf{z}_m^{(\mathrm{s})} \middle| \mathbf{y}_{1:m}^{(\mathrm{s})}, \zeta_m = k\right)$ gemäß

       $$\hat{\mathbf{z}}_{m|m,k}^{(\mathrm{s})} = \sum_{i=1}^{I} P_{m,k,i}^{(\mathrm{FUS})} \hat{\mathbf{z}}_{m|m,i,k}^{(\mathrm{s})}, \tag{5.245}$$

       $$\hat{\mathbf{\Sigma}}_{\mathbf{z}_{m|m,k}^{(\mathrm{s})}} = \sum_{i=1}^{I} P_{m,k,i}^{(\mathrm{FUS})} \left[\hat{\mathbf{\Sigma}}_{\mathbf{z}_{m|m,i,k}^{(\mathrm{s})}} + \left(\hat{\mathbf{z}}_{m|m,i,k}^{(\mathrm{s})} - \hat{\mathbf{z}}_{m|m,k}^{(\mathrm{s})}\right)\left(\hat{\mathbf{z}}_{m|m,i,k}^{(\mathrm{s})} - \hat{\mathbf{z}}_{m|m,k}^{(\mathrm{s})}\right)^T\right]. \tag{5.246}$$

    3. Berechne für $i \in \{1,\ldots,I\}$ die A-posteriori-Modellwahrscheinlichkeiten gemäß

       $$P_{m|m,i} \propto \sum_{k=1}^{I} P_{m,k,i}^{(\mathrm{FUS})}. \tag{5.247}$$

    4. Berechne den Mittelwertvektor und die Kovarianzmatrix der A-posteriori-Verteilungsdichtefunktion $p\left(\mathbf{z}_m^{(\mathrm{s})} \middle| \mathbf{y}_{1:m}^{(\mathrm{s})}\right)$ gemäß

       $$\hat{\mathbf{z}}_{m|m}^{(\mathrm{s})} = \sum_{i=1}^{I} P_{m|m,i} \hat{\mathbf{z}}_{m|m,i}^{(\mathrm{s})}, \tag{5.248}$$

       $$\hat{\mathbf{\Sigma}}_{\mathbf{z}_{m|m}^{(\mathrm{s})}} = \sum_{i=1}^{I} P_{m|m,i} \left[\hat{\mathbf{\Sigma}}_{\mathbf{z}_{m|m,i}^{(\mathrm{s})}} + \left(\hat{\mathbf{z}}_{m|m,i}^{(\mathrm{s})} - \hat{\mathbf{z}}_{m|m}^{(\mathrm{s})}\right)\left(\hat{\mathbf{z}}_{m|m,i}^{(\mathrm{s})} - \hat{\mathbf{z}}_{m|m}^{(\mathrm{s})}\right)^T\right]. \tag{5.249}$$

  - **Extraktion der Schätzungen**:
    1. Extrahiere $\hat{\mathbf{n}}_m^{(\mathrm{s})}$ und $\hat{\mathbf{\Sigma}}_{\check{\mathbf{n}}_m^{(\mathrm{s})}}$ aus $\hat{\mathbf{z}}_{m|m}^{(\mathrm{s})}$ und $\hat{\mathbf{\Sigma}}_{\mathbf{z}_{m|m}^{(\mathrm{s})}}$ gemäß (5.235).
    2. **Wenn** $m \geq L_\mathrm{C}$ **dann**
       - Extrahiere $\hat{\mathbf{x}}_{m-L_\mathrm{C}+1}^{(\mathrm{s})}$ und $\hat{\mathbf{\Sigma}}_{\check{\mathbf{x}}_{m-L_\mathrm{C}+1}^{(\mathrm{s})}}$ aus $\hat{\mathbf{z}}_{m|m}^{(\mathrm{s})}$ und $\hat{\mathbf{\Sigma}}_{\mathbf{z}_{m|m}^{(\mathrm{s})}}$ gemäß (5.236).

       Ende wenn

  Ende für

6. Experimentelle Untersuchungen

In diesem Kapitel wird das zuvor vorgestellte Verfahren zur BAYES'schen Merkmalsverbesserung ausführlich experimentell untersucht. Die dazu verwendeten Sprachdatenbanken werden zunächst in Kap. 6.1 im Detail beschrieben. Anschließend werden in Kap. 6.2 Erkennungsergebnisse derzeit existierender Referenzverfahren auf diesen Datenbanken präsentiert, um die Schwierigkeit der Spracherkennung unter Präsenz von Nachhall und Störungen vor Augen zu führen. Danach werden in Kap. 6.4 Resultate von Voruntersuchungen zur Merkmalsverbesserung dargelegt. In Kap. 6.5 folgen experimentelle Ergebnisse zur reinen Merkmalsenthallung, wobei besonderes Augenmerk auf den Einfluss des A-priori-Modells und des Beobachtungsmodells auf die Leistungsfähigkeit des Verfahrens gelegt wird. Abschließend werden in Kap. 6.6 experimentelle Ergebnisse zur gemeinsamen Merkmalsenthallung und -entstörung dargeboten.

6.1. Sprachdatenbanken und Konfigurationen der Spracherkenner

Die Datenbanken wurden derart ausgewählt bzw. selbst modifiziert, dass die Leistungsfähigkeit des zuvor vorgestellten Verfahrens zur Merkmalsverbesserung sowohl für Spracherkennungsaufgaben mit kleinem als auch großem Vokabular unter Einfluss von Nachhall und Hintergrundstörungen untersucht werden konnte. Als Aufgabe mit einem kleinem Vokabular wurde eine Erkennung von Ziffernketten betrachtet. Zu diesem Zweck wurde die AURORA5-Datenbank verwendet, die in Kap. 6.1.1 beschrieben wird.

Soweit es dem Autor bekannt ist, existiert bislang keine Sprachdatenbank mit großem Vokabular, bei der die Sprachäußerungen in halligen Umgebungen aufgenommen worden sind. Aus diesem Grund wurde für die Erkennungsaufgabe mit großem Vokabular die AURORA4-Datenbank, die nur durch Hintergrundstörungen beeinflusste Sprachäußerungen beinhaltet, herangezogen und geeignet modifiziert, um zusätzlich den Effekt des Nachhalls einzubeziehen. Die AURORA4-Datenbank sowie die daran vorgenommenen Modifikationen sind in Kap. 6.1.2 dokumentiert.

6.1.1. AURORA5-Datenbank

Die AURORA5-Datenbank [Hir07] wurde vorwiegend zur Untersuchung der Leistungsfähigkeit von Spracherkennungssystemen im Freihandsprachbetrieb in Gegenwart von Hintergrundstörungen entwickelt. Sie besteht aus Sprachäußerungen erwachsener Personen von Ziffernketten in amerikanischem Englisch und basiert auf der ***T**exas **I**nstruments* (*TI*)-Digits-Datenbank. Die für die *TI*-Digits-Datenbankmit mit einer Abtastrate von 20 kHz aufgenom-

menen Sprachsignale wurden dabei für die Erzeugung der AURORA5-Datenbank mit 8 kHz unterabgetastet. Das Vokabular besteht aus insgesamt 11 Wörtern, da die Ziffer Null in den beiden englischen Aussprachevarianten *zero* und *oh* vorkommt.

Das Hauptaugenmerk der Ersteller der AURORA5-Datenbank lag auf der Betrachtung von realistischen Anwendungsszenarien, von denen zwei besondere ausgewählt wurden. Diese umfassen erstens eine Freisprechsituation innerhalb eines Fahrzeugs unter Präsenz von Hintergrundstörungen, bei der beispielsweise Geräte von einer Person innerhalb des Fahrzeugs bedient oder Informationen von einem entfernten Sprachserver über das Telefon abgerufen werden, und zweitens eine Freisprechsituation innerhalb eines Büros oder Wohnzimmers, bei der beispielsweise ein Telefon oder Audio- und Videogeräte von einer Person bedient werden. In dieser Arbeit wird nur auf denjenigen Teil der Datenbank Bezug genommen, der das zweite Anwendungsszenario betrifft, da der Einfluss des Nachhalls, der hier im Vordergrund steht, innerhalb von Räumen deutlich größer als innerhalb von Fahrzeugen ist.

Zur Erstellung der Datenbank wurden die Sprachsignale nicht tatsächlich mit Freisprechmikrophonen aufgenommen, sondern vielmehr künstlich durch eine Faltung von sauberen Sprachsignalen mit zeitinvarianten RIAs berechnet. Die RIAs wurden mit Hilfe der Spiegelquellenmethode [All79] erzeugt, wobei zusätzlich später Nachhall zum Zweck eines natürlichen Nachhallklanges hinzugefügt wurde. Für die Spiegelquellenmethode wurden zwei virtuelle Räume, bezeichnet als Büro und Wohnzimmer, angenommen, wobei für jeden der beiden virtuellen Räume drei unterschiedliche Versionen von RIAs berechnet wurden. Diese unterschieden sich vorwiegend in der simulierten Nachhallzeit, die für das Büro jeweils etwa 0,3 s, 0,35 s und 0,4 s und für das Wohnzimmer etwa 0,4 s, 0,45 s und 0,5 s betrug. Die Werte des *DRR* liegen bei allen RIAs im Bereich zwischen -5 dB und -7 dB. Zur künstlichen raumspezifischen Verhallung jeder einzelnen Sprachäußerung wurde jeweils eine der drei betreffenden RIAs zufällig ausgewählt. Eine detaillierte Darstellung aller verwendeten RIAs samt ihren log-MEL-spektralen Repräsentationen findet sich in Kap. A.3 im Anhang.

Als Trainingsdaten werden in dieser Arbeit stets nur die in dem Trainingsdatensatz der AURORA5-Datenbank enthaltenen 8623 Sprachäußerungen in Form von sauberen Signalen verwendet. Die Testdaten bestehen aus 8700 Sprachäußerungen mit insgesamt 28 583 Wörtern. Neben den sauberen Sprachsignalen liegen verhallte Versionen derselben Signale für die beiden simulierten Räume vor, die keine Hintergrundstörungen beinhalten. Weiterhin enthalten die Testdaten gemeinsam gestörte und verhallte Versionen derselben Sprachsignale, die durch additive Überlagerung der verhallten Sprachsignale mit Störsignalen mit einem *SNR* zwischen 0 dB und 15 dB erzeugt wurden. Als Störsignale wurden zufällige Ausschnitte aus 5 Signalen der Länge von jeweils etwa 3 Minuten herangezogen, welche in einem Einkaufszentrum, einem Restaurant, einer Ausstellungshalle, einem Büro und einer Hotelempfangshalle aufgenommen wurden.

Für den Spracherkenner wurde ein Unigramm als Sprachmodell und ein HMM-basiertes akustisches Modell verwendet. Für jedes der 11 Wörter wurde ein geschlechtsunabhängiges HMM mit Links-Rechts-Topologie bestehend aus insgesamt 16 Zuständen verwendet, wobei das Überspringen von Zuständen nicht zugelassen war. Die Emissionsverteilungsdichtefunktionen für jeden dieser Zustände wurden durch ein *GMM* mit 4 Mischungskomponenten beschrieben. Außerdem wurde ein HMM zur Modellierung von Sprachpausen bestehend aus 3 Zuständen eingeführt, wobei ebenfalls ein *GMM* mit 4 Mischungskomponenten zur Darstellung der Emissionsverteilungsdichtefunktionen genutzt wurde. Für die GAUSS-Mischungsverteilungen wurden diagonale Kovarianzmatrizen zugrunde gelegt. Der

Spracherkenner wurde mit Hilfe von *HTK* [YEG+06] in einem überwachten Modus trainiert, wobei zwar die Transkription der Sprachäußerungen bekannt war, jedoch nicht die zeitliche Anpassung der Transkription an die Äußerung. Die Merkmalsextraktion wurde wie in Kap. 2.1 beschrieben mit Hilfe des *ETSI-SFE* durchgeführt, so dass als Merkmale die *MFCC*s gemeinsam mit den DELTA- und DELTA-DELTA-Merkmalen (siehe (2.9)) dienten.

6.1.2. Modifizierte AURORA4-Datenbank

Die AURORA4-Datenbank [PP02] wurde unter anderem mit dem Ziel entwickelt, die Robustheit von Spracherkennungssystemen mit unterschiedlichen Verfahren zur Merkmalsextraktion gegenüber additiven Störungen sowie der Variation von Mikrophoncharakteristiken zu untersuchen. Sie besteht aus Aufnahmen von kontinuierlich gesprochener englischer Sprache mit einem Vokabular von 5000 Wörtern basierend auf dem sogenannten ***D**efense **A**dvanced **R**esearch **P**rojects **A**gency* (*DARPA*) ***W**all **S**treet **J**ournal* (*WSJ*) *Corpus* [PB92], wobei die Grundlage für die Äußerungen gelesene Zeitungsartikel aus dem *WSJ* bilden. Für die Experimente in dieser Arbeit wurden die unterabgetasteten Versionen der Sprachsignale verwendet, wobei die Abtastrate 8 kHz betrug. Die Sprachsignale sind gemäß dem G.712 Standard der Internationalen Fernmeldeunion (engl. ***I**nternational **T**elecommunication **U**nion* (*ITU*)) [Int96] gefiltert.

Die Trainingsdaten beinhalten unter anderem Sprachäußerungen in Form von sauberen Signalen bestehend aus 7138 Sätzen von insgesamt 83 verschiedenen Sprechern und besitzen eine Aufnahmedauer von etwa 14 Stunden. Für alle Experimente bezüglich der AURORA4-Datenbank wurden ausschließlich diese Daten zum Training des Erkenners verwendet. Als Testdatensatz wurde der sogenannte ***N**ational **I**nstitute of **S**tandards and **T**echnology* (*NIST*) Nov'92 Evaluierungsdatensatz betrachtet. Dieser umfasst in seiner originalen Form insgesamt 14 Testsätze, von denen 7 mit einem Sennheiser HMD414 Mikrophon und 7 mit 18 weiteren Mikrophonen aufgenommen worden sind. In dieser Arbeit wurden lediglich die mit dem Sennheiser HMD414 Mikrophon gemachten Aufnahmen herangezogen. Die 7 Testsätze stellen jeweils 7 unterschiedliche Versionen eines Datensatzes bestehend aus 166 Sätzen und 2715 Wörtern dar. Eine dieser Versionen bilden die sauberen Sprachsignale, während die weiteren 6 Versionen durch additive Überlagerung der sauberen Sprachsignale mit unterschiedlichen Arten von Störsignalen mit einem *SNR* zwischen 5 dB und 15 dB entstanden sind. Die Störsignale sind unter anderem innerhalb von Fahrzeugen oder auf der Straße aufgenommen worden und sind daher im Hinblick auf die Untersuchungen dieser Arbeit ungeeignet, da sie für Innenräume untypisch sind. Da zudem bei der Erstellung der AURORA4-Datenbank keine Berücksichtigung von Freisprechszenarien stattfand, wurden diese 6 Testsätze hier vollständig verworfen. Statt dessen wurde durch den Autor ein modifizierter Testdatensatz unter Einbezug von Nachhall und typischen Störungen aus Innenräumen erstellt. Dazu wurden die sauberen Sprachsignale des Standardtestdatensatzes der AURORA4-Datenbank mit denselben künstlich erzeugten Raumimpulsantworten wie bei der AURORA5-Datenbank gefaltet, um verhallte Testsprachsignale für die zwei virtuelle Räume, bezeichnet als Büro und Wohnzimmer, zu erhalten. Zusätzlich wurden die verhallten Sprachsignale additiv mit Störungen mit einem *SNR*s von 0 dB, 5 dB, 10 dB und 15 dB überlagert, um gemeinsam verhallte und gestörte Sprachsignale zu erzeugen.

Für den Spracherkenner wurde ein Bigramm als Sprachmodell und ein HMM-basiertes akustisches Modell verwendet. Im Gegensatz zur AURORA5-Datenbank wurden hierbei

HMMs für einzelne Triphone trainiert, wobei das gesamte akustische Modell etwa 3240 Zustände aufweist, deren Emissionsverteilungsdichtefunktionen durch *GMM*s mit jeweils 10 Mischungskomponenten, gekennzeichnet durch diagonale Kovarianzmatrizen, dargestellt wurden. Das Training des Erkenners fand mit Hilfe von *HTK* in einem überwachten Modus statt. Um dem Sprachmodell gegenüber dem akustischen Modell mehr Gewicht bei der Decodierung zu verleihen, wurde die Konstante $\alpha^{(\mathrm{SM})}$ zu 16 gesetzt. Die Merkmalsextraktion erfolgte wie auch bei der AURORA5-Datenbank mit dem *ETSI-SFE* gemäß der Beschreibung in Kap. 2.1.

6.2. Referenzergebnisse

Als Qualitätsmaß zur Bewertung der Leistungsfähigkeit eines Systems zur automatischen Spracherkennung fungiert in dieser Arbeit ausschließlich die erzielte Wortfehlerrate λ_w, welche durch

$$\lambda_w := \frac{N_{\mathrm{Subst}} + N_{\mathrm{Ausl}} + N_{\mathrm{Einf}}}{N_{\mathrm{Ges}}} \tag{6.1}$$

definiert ist. Dabei bezeichnen N_{Subst}, N_{Ausl}, N_{Einf} und N_{Ges} in dieser Reihenfolge jeweils die Anzahl der fälschlicherweise ersetzten, ausgelöschten und eingefügten Wörter sowie die Gesamtanzahl der Wörter innerhalb der Testdaten.

Die Referenzergebnisse, welche ohne Anwendung jeglicher Merkmalsverbesserung für die AURORA5-Datenbank bzw. die modifizierte AURORA4-Datenbank erzielt wurden, sind in Tab. 6.1 bzw. Tab. 6.2 aufgeführt. Für die modifizierte AURORA4-Datenbank ist wie in

Tabelle 6.1.: *Wortfehlerraten λ_w [%] für die AURORA5-Datenbank erzielt mit dem ETSI-SFE.*

		Raum	
		Büro	Wohnzimmer
SNR [dB]	∞	6,32	14,94
	15	19,93	35,58
	10	44,75	57,38
	5	71,73	79,01
	0	88,10	89,72

der Literatur üblich die Wortfehlerrate zusätzlich in die Raten der Ersetzungs- der Auslöschungs- und Einfügefehler definiert durch

$$\lambda_{\mathrm{Subst}} := \frac{N_{\mathrm{Subst}}}{N_{\mathrm{Ges}}}, \quad \lambda_{\mathrm{Ausl}} := \frac{N_{\mathrm{Ausl}}}{N_{\mathrm{Ges}}}, \quad \lambda_{\mathrm{Einf}} := \frac{N_{\mathrm{Einf}}}{N_{\mathrm{Ges}}} \tag{6.2}$$

aufgeschlüsselt. Die einzelnen Fehlerraten wurden aus den erkannten Wortsequenzen mit Hilfe von *HTK* [YEG+06] berechnet. Für die sauberen Testsprachsignale liegt die Wortfehlerrate für die AURORA5-Datenbank bei 0,66 % und für die modifizierte AURORA4-Datenbank bei 14,00 %. Die Referenzergebnisse zeigen unter anderem den starken negativen Einfluss des Nachhalls auf die Wortfehlerrate, die sich beispielsweise für das Wohnzimmerszenario um etwa 2200 % für die AURORA5-Datenbank und um etwa 500 % für die

Tabelle 6.2.: *Fehlerraten [%] für die modifizierte AURORA4-Datenbank erzielt mit dem ETSI-SFE.*

		Raum							
		Büro				Wohnzimmer			
		λ_{Subst}	λ_{Ausl}	λ_{Einf}	λ_w	λ_{Subst}	λ_{Ausl}	λ_{Einf}	λ_w
	∞	34,84	5,52	7,00	**47,37**	56,06	10,87	6,52	**73,44**
	15	49,13	9,80	8,58	**67,51**	58,64	16,80	7,55	**82,98**
SNR [dB]	10	57,90	20,77	7,11	**85,78**	55,58	32,15	4,05	**91,79**
	5	47,88	44,27	2,39	**94,55**	38,93	56,98	1,25	**97,16**
	0	27,55	70,72	0,66	**98,93**	18,97	80,00	0,15	**99,12**

modifizierte AURORA4-Datenbank relativ im Vergleich zu den sauberen Testsprachsignalen erhöht. Weiterhin wächst die Wortfehlerrate bei vorhandener Störung mit abnehmendem *SNR* und erreicht bei einem *SNR* von 0 dB für beide Datenbanken Werte über 85 %, welche für praktische Anwendungen nicht mehr akzeptabel sind. Diese Ergebnisse verdeutlichen den Bedarf an Verfahren zur robusten Spracherkennung in Gegenwart von Nachhall und Hintergrundstörungen.

6.3. Ergebnisse alternativer Verfahren

Zusätzlich zu den Ergebnissen des Standarderkennungssystems werden in diesem Abschnitt die Resultate dreier alternativer Referenzverfahren präsentiert.

Das erste Verfahren ist dadurch gekennzeichnet, dass die Merkmalsextraktion mit dem sogenannten *ETSI-**A**dvanced **F**ront **E**nd* (*AFE*) [ETSa] bewerkstelligt wird. Dieses wurde speziell für den Zweck einer störungsrobusten Spracherkennung entwickelt und bietet eine sehr hohe Leistungsfähigkeit, die bis heute kaum von einem anderen Verfahren überboten wird. Es unterscheidet sich vom Standardverfahren des *ETSI-SFE* im Wesentlichen durch ein zusätzliches zweistufiges WIENER-Filter zur Störsignalunterdrückung sowie eine Blindentzerrung (engl. *blind equalization*) zur Kompensation einer akustischen Fehlanpassung, welche durch die Verwendung unterschiedlicher Aufnahmegeräte beim Training und beim Test entsteht.

Das zweite Verfahren nutzt weiterhin das *ETSI-SFE* zur Merkmalsextraktion. Für das Training des Spracherkenners wurden jedoch nicht wie gewöhnlich die sauberen Trainingssprachsignale verwendet. Statt dessen wurde der Erkenner separat für jedes Testszenario, d.h. für das Büro und das Wohnzimmer, mit raumspezifischen verhallten Sprachsignalen trainiert. Die RIAs zur Berechnung der verhallten Trainingsignale wurden mit Hilfe der Spiegelquellenmethode [All79] künstlich erzeugt (siehe Kap. 6.4). Ein derartig grob auf das Testszenario abgestimmtes Training ist für die Praxis durchaus geeignet.

Beim dritten Verfahren erfolgt eine Adaption der HMM-Parameter auf den Effekt des Nachhalls und der Hintergrundstörungen gemäß der *PMC*-Methode [HF08]. Dabei wird zur Anpassung der Emissionsverteilungsdichtefunktionen einzelner HMM-Zustände der Einfluss vorhergehender HMM-Zustände über ein deterministisches Modell der *EDC* berücksichtigt. Insbesondere soll darauf hingewiesen werden, dass im Hinblick auf die Adaption auf Hintergrundstörungen für jede Testsprachäußerung das Störsignal als instationär ange-

nommen wird. Deshalb wird zunächst dessen zeitvariante Charakteristik mit Hilfe einer *VAD* aus dem verhallten und gestörten Sprachsignal gemäß [HE95] geschätzt. Anschließend erfolgt eine entsprechende dynamische Adaption der HMM-Parameter.

Die Ergebnisse der drei Referenzverfahren für die AURORA5-Datenbank sind in Tab. 6.3 dargestellt. Dabei wurden die Resultate, welche sich auf die Adaption der HMM-Parameter

Tabelle 6.3.: *Wortfehlerraten λ_w [%] für die AURORA5-Datenbank erzielt mit alternativen Verfahren.*

		Raum	
		Büro	Wohnzimmer
SNR [dB]	∞	6,11	14,53
	15	10,92	21,31
	10	17,26	29,17
	5	30,09	43,06
	0	51,41	62,65

(a) *ETSI-AFE*

		Szenario	
		Büro	Wohnzimmer
SNR [dB]	∞	1,29	2,61
	15	15,44	14,58
	10	38,31	51,19
	5	67,81	77,88
	0	87,63	91,88

(b) Training des Erkenners mit verhallten Sprachsignalen

		Szenario	
		Büro	Wohnzimmer
SNR [dB]	∞	3,30	8,00
	15	6,20	9,20
	10	11,50	16,90
	5	24,30	32,00
	0	49,20	60,00

(c) Adaption der HMM-Parameter gemäß der *PMC*-Methode (Ergebnisse aus [HF08])

beziehen, direkt aus Diagrammen in [HF08] abgelesen. Es muss jedoch bei deren Beurteilung darauf geachtet werden, dass zu ihrer Erzeugung eine geringfügig abweichende Konfiguration des Merkmalsextraktors und des Spracherkenners verwendet worden ist, so dass streng genommen keine direkte Vergleichbarkeit gewährleistet ist. So wurde einerseits zur Berechnung der dynamischen Merkmale $\Delta y^{(c)}_{m,\kappa'}$ das entsprechende Zeitfenster kleiner als in den Experimenten in dieser Arbeit gewählt, wobei die Konstante I_1 zu 3 anstatt 4 gesetzt wurde (siehe Tab. 2.1). Andererseits wurden an Stelle von geschlechtsunabhängigen geschlechtsspezifische HMMs verwendet, wobei die Emissionsverteilungsdichtefunktionen der HMM-Zustände einzelner Wörter durch GAUSS-Mischungsverteilungsdichtefunktionen mit jeweils 2 Komponenten modelliert wurden. Zur Beschreibung der Emissionsverteilungsdichtefunktionen der Zustände des Sprachpause-HMM wurden 8 GAUSS-förmige Mischungskomponenten eingesetzt. Es ist jedoch davon auszugehen, dass die genannten Abweichungen der Spracherkennerkonfiguration nur geringfügige Auswirkungen auf die Leistungsfähigkeit des Spracherkenners ausüben, so dass zumindest ein grober Vergleich zulässig ist.

Weiterhin sind in Tab. 6.4 die Ergebnisse von zwei Referenzverfahren für die modifizierte AURORA4-Datenbank aufgeführt. Bedauerlicherweise existieren in [HF08] keine detail-

Tabelle 6.4.: *Fehlerraten [%] für die modifizierte AURORA4-Datenbank erzielt mit alternativen Verfahren.*

		Raum							
		Büro				Wohnzimmer			
		λ_{Subst}	λ_{Ausl}	λ_{Einf}	λ_w	λ_{Subst}	λ_{Ausl}	λ_{Einf}	λ_w
SNR [dB]	∞	34,03	6,48	6,08	**46,59**	55,99	11,16	5,45	**72,60**
	15	35,14	7,07	7,40	**49,61**	50,68	8,14	9,10	**67,92**
	10	44,01	9,47	9,43	**62,91**	59,08	10,72	9,21	**79,01**
	5	54,84	13,41	8,73	**76,98**	67,33	15,29	7,00	**89,61**
	0	63,98	21,62	6,11	**91,71**	67,18	24,71	3,54	**95.43**

(a) *ETSI-AFE*

		Raum							
		Büro				Wohnzimmer			
		λ_{Subst}	λ_{Ausl}	λ_{Einf}	λ_w	λ_{Subst}	λ_{Ausl}	λ_{Einf}	λ_w
SNR [dB]	∞	18,01	3,06	3,17	**24,24**	26,26	6,08	3,98	**36,32**
	15	28,40	4,90	12,97	**46,26**	36,13	6,85	11,57	**54,55**
	10	43,09	9,54	13,33	**65,97**	48,25	12,78	10,20	**71,23**
	5	51,90	24,83	7,07	**83,79**	47,73	36,24	4,68	**88,66**
	0	36,39	56,35	1,92	**94,66**	28,73	66,52	1,62	**96,87**

(b) Training des Erkenners mit verhallten Sprachsignalen

lierten Ergebnisse für diese Datenbank. Jedoch haben die Autoren von [HF08] ein ähnliches Experiment durchgeführt, wobei eine triphonbasierte HMM-Adaption auf den Nachhall in einem Büro mit einer Nachhallzeit T_{60} von etwa 0,4 s vorgenommen wurde. Die Wortfehlerrate von 48,8 %, welche mit den auf sauberen Sprachsignalen trainierten HMMs auf den verhallten Testsprachsignalen erzielt wurde, konnte unter Verwendung der HMM-Adaption auf 39,8 % reduziert werden. Für die Erkennung von Sprachäußerungen in Form von sauberen Sprachsignalen wird eine Wortfehlerrate von 11,21 % angegeben. Obwohl in diesem Experiment die Abtastfrequenz des Sprachsignals f_{A} 16 kHz beträgt sowie die Art der Merkmalsextraktion und die Konfiguration des Erkenners geringfügig von der in dieser Arbeit verwendeten abweicht (vgl. Kap. 6.1.2 mit [HF08]), lässt die Ähnlichkeit der Worterkennungsraten für den Fall ohne Adaption (vgl. 48,8 % mit 47,37 % aus Tab. 6.2) eine gewisse Vergleichbarkeit zu.

Im Hinblick auf die Interpretation der Ergebnisse in Tab. 6.3a und Tab. 6.4a lässt sich zunächst feststellen, dass das *ETSI-AFE* generell nicht dazu geeignet ist, die Wortfehlerraten in störungsfreien halligen Umgebungen gegenüber dem *ETSI-SFE* zu verbessern. Dieses kann darauf zurückgeführt werden, dass für die Berechnung der Übertragungsfunktion des WIENER-Filters keine Berücksichtigung der Korrelation zwischen dem Direktanteil und den durch den Nachhall bedingten Anteil der Sprache stattfindet. Bei Vorhandensein von zusätzlicher unkorrelierter, additiver Störung lässt sich dann wiederum wie erwartet eine deutliche

Leistungsverbesserung gegenüber dem *ETSI-SFE* feststellen.

Die Ergebnisse für das Training des Erkenners mit künstlich verhallten Sprachsignalen in Tab. 6.3b und Tab. 6.4b zeigen ein gegensätzliches Verhalten. Während für störungsfreie hallige Umgebungen für beide Datenbanken ein deutliches Absinken der Wortfehlerrate gegenüber dem Standardtraining zu verzeichnen ist, nahm die Leistungsfähigkeit bei Vorhandensein von zusätzlicher additiver Störung mit sinkendem *SNR* ab. Diese Resultate sind nicht überraschend, da bei der Erkennung additive Störungen vollkommen außer Betracht gelassen wurden.

Die Adaption der HMM-Parameter führte in Abwesenheit von Hintergrundstörungen zu einer beeindruckenden Reduktion der Wortfehlerrate, wobei auf der AURORA5-Datenbank für beide Räume etwa 50 % der durch den Nachhall verursachten Fehler korrigiert werden konnten. Im Vergleich dazu betrug der Anteil der korrigierten Fehler auf der modifizierten AURORA4-Datenbank für das Büro nur noch etwa 24 %. Die Leistungsfähigkeit wie beim Training des Erkenners mit künstlich verhallten Sprachsignalen konnte jedoch auf beiden Datenbanken nicht erreicht werden, was zum Teil sicherlich darauf zurückzuführen ist, dass der linksseitige Kontext bei der Adaption der HMMs nicht hinreichend genug berücksichtigt wurde.

In Gegenwart von Hintergrundstörungen liefert die Modelladaption die besten Ergebnisse im Vergleich mit den beiden anderen vorgestellten Verfahren. Ein wesentlicher Aspekt dabei ist höchstwahrscheinlich die dynamische Adaption der Charakteristik der Hintergrundstörung.

6.4. Voruntersuchungen zum Beobachtungsmodell

Die tatsächliche praktische Durchführung der Merkmalsverbesserung gemäß der Beschreibung in Kap. 5.3 erfordert vorab die Festlegung oder Bestimmung gewisser Parameter. Dazu gehören unter anderem Schätzungen der Koeffizienten der RIA im log-MEL-spektralen Bereich sowie die Mittelwerte und Kovarianzmatrizen des Beobachtungsfehlers.

Für die weiteren Untersuchungen wurde von einer optimalen mittleren geschätzten Nachhallzeit $\hat{T}_{60}$ von 0,35 s für das Büro und 0,45 s für das Wohnzimmer ausgegangen, welche unter Zuhilfenahme von (5.143) in geschätzten Abklingkonstanten $\hat{\tau}_h$ von etwa $7,05 \cdot 10^{-2}$ bzw. $6,22 \cdot 10^{-2}$ resultierten. Weiterhin ist zu berücksichtigen, dass für jede der beiden untersuchten Datenbanken die Trainings- und Testdaten derselben Energienormierung unterliegen, so dass der Skalierungsparameter für die RIA $\hat{\sigma}_h$ gemäß (5.171) gewählt wurde. Eine sinnvolle Festlegung der RIA-Länge $\hat{L}_h$, aus der sich anschließend die gesuchte Größe $\hat{L}_H$ gemäß (5.112) berechnen lässt, kann durch die Festlegung des Parameters ε_h gemäß (5.157) geschehen. Sie ist jedoch nur bei der Verwendung des nichtrekursiven Beobachtungsmodells notwendig, dessen Modellierungsfähigkeit im Allgemeinen durch eine Verringerung des Wertes von ε_h verbessert wird. Dabei ist jedoch zu beachten, dass beim Unterschreiten eines gewissen Wertebereiches aufgrund von Modellunzulänglichkeiten und Parameterfehlschätzungen keine genauere Modellierung mehr zu erwarten ist.

Beruhend auf diesen Überlegungen wurde ein sinnvoller Wert von ε_h experimentell bestimmt. Dazu wurde mit Hilfe von Trainingsdaten der Beobachtungsfehler jeweils gemäß (5.192) für unterschiedliche Werte von ε_h berechnet. Dieses wurde zunächst nur für den störungsfreien Fall umgesetzt. Um eine praxisrelevante Situation zu simulieren, in der ge-

wöhnlich keine verhallten Sprachsignale für das Erkennungsszenario zur Verfügung stehen, wurden die verhallten Signale durch Faltung von sauberen Sprachsignalen mit durch die Spiegelquellenmethode [All79] künstlich berechneten RIAs generiert. Für die Spiegelquellenmethode wurde für beide Szenarien, d.h. für das Büro und das Wohnzimmer, derselbe quaderförmige Raum gemäß Abb. 6.1 eingesetzt. Die Ausmaße des Raumes wurden basierend auf der Annahme, dass die Raumgröße in praktischen Anwendungen in der Regel vorab unbekannt ist, vollkommen willkürlich gewählt. Der Beobachtungsfehler basierte für beide

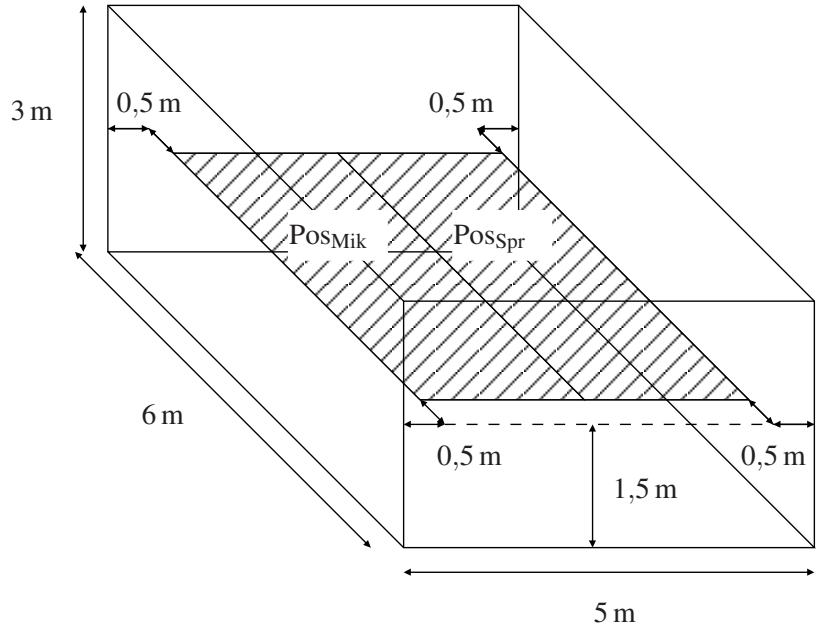


Abbildung 6.1.: *Zur Anwendung der Spiegelquellenmethode verwendeter quaderförmiger, virtueller Raum, in dem die Position des Sprechers und des Mikrophons gleichverteilt innerhalb der durch die Pos_{Spr} und Pos_{Mik} gekennzeichneten Flächen variiert wurde.*

Datenbanken auf jeweils 575 Sprachäußerungen, wobei für jede einzelne Äußerung zufällig eine von 50 individuellen RIAs verwendet wurde, für deren Erzeugung die Position des Sprechers und des Mikrophons zufällig innerhalb der in Abb. 6.1 durch $\mathrm{Pos}_{\mathrm{Spr}}$ und $\mathrm{Pos}_{\mathrm{Mik}}$ gekennzeichneten Flächen ausgewählt wurde. Zusätzlich wurde für die Berechnung jeder einzelnen RIA die Nachhallzeit T_{60} gleichförmig zufällig aus dem Intervall $[0{,}3\,\mathrm{s}, 0{,}4\,\mathrm{s}]$ für das Büro und aus dem Intervall $[0{,}4\,\mathrm{s}, 0{,}5\,\mathrm{s}]$ für das Wohnzimmer selektiert.

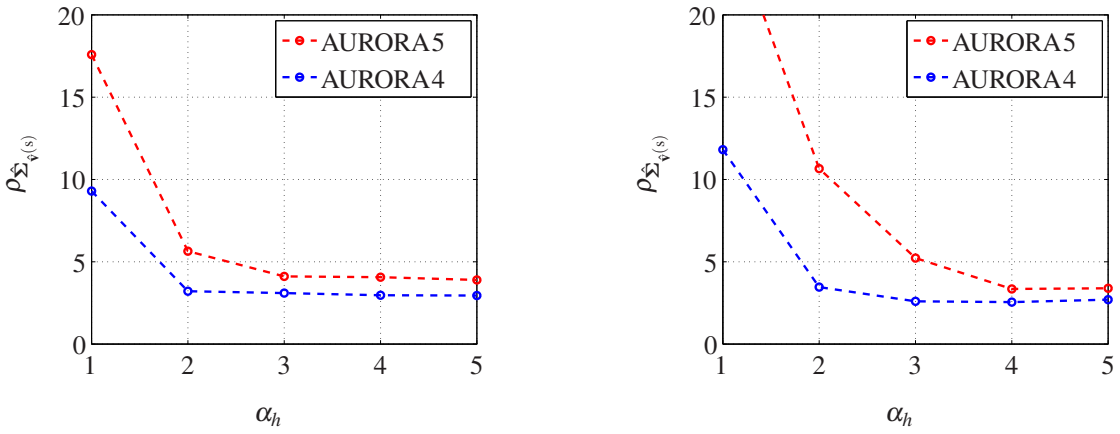
In einem ersten Experiment wurde für unterschiedliche Werte von ε_h eine Schätzung $\hat{\Sigma}_{\hat{\mathbf{v}}(\mathrm{s})}$ für die Kovarianzmatrix $\Sigma_{\hat{\mathbf{v}}(\mathrm{s})}$ empirisch mit der *Maximum-Likelihood*-Methode aus der Folge der Beobachtungsfehler bestimmt und anschließend ihr Spektralradius

$$\rho_{\hat{\Sigma}_{\hat{\mathbf{v}}(\mathrm{s})}} := \max\left\{ |\lambda| \;\middle|\; \lambda \text{ ist Eigenwert von } \hat{\Sigma}_{\hat{\mathbf{v}}(\mathrm{s})} \right\} \tag{6.3}$$

berechnet. Der Spektralradius $\rho_{\hat{\Sigma}_{\hat{\mathbf{v}}(\mathrm{s})}}$ diente dabei als Maß für die im Beobachtungsmodell enthaltene Unsicherheit. Die resultierenden Werte in Abhängigkeit des negativen Exponenten von ε_h zur Basis 10, definiert durch

$$\alpha_h := -\log_{10}(\varepsilon_h), \tag{6.4}$$

sind für die AURORA5- und die modifizierte AURORA4-Datenbank in Abb. 6.2 dargestellt. Es lässt sich erkennen, dass zunächst für beide Datenbanken der Spektralradius $\rho_{\hat{\Sigma}_{\hat{\nu}(s)}}$ mit

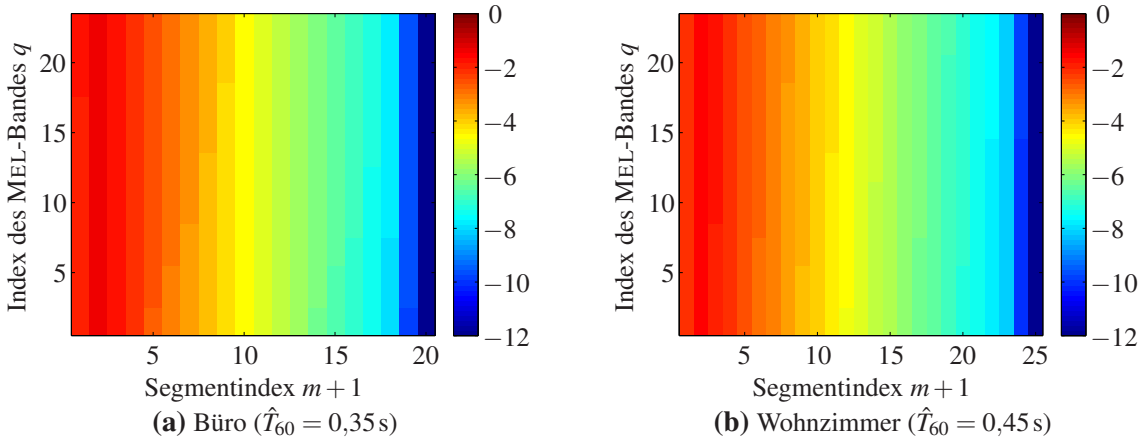


(a) Büro **(b)** Wohnzimmer

Abbildung 6.2.: *Spektralradius $\rho_{\hat{\Sigma}_{\hat{\nu}(s)}}$ der empirisch berechneten Kovarianzmatrix des Beobachtungsfehlers $\hat{\Sigma}_{\hat{\nu}(s)}$ in Abhängigkeit von α_h.*

wachsenden Werten von α_h abnimmt, wobei ab etwa einem Wert von $\alpha_h = 3$ keine oder nur noch eine relativ marginale Verringerung des Spektralradius auftritt. Als Kompromiss wurde deshalb als Grundlage für alle weiteren Experimente im Zusammenhang mit der nichtrekursiven Beobachtungsfunktion $\varepsilon_h = 10^{-3}$ angenommen, was zu approximativen Längen der Repräsentation der RIA im log-MEL-spektralen Bereich von $\hat{L}_H = 19$ für das Büro und $\hat{L}_H = 24$ für das Wohnzimmer führte.

Die entsprechenden Approximationen der log-MEL-spektralen Repräsentationen der RIAs beider Räume werden in Abb. 6.3 veranschaulicht. Einen qualitativen Eindruck der Güte der

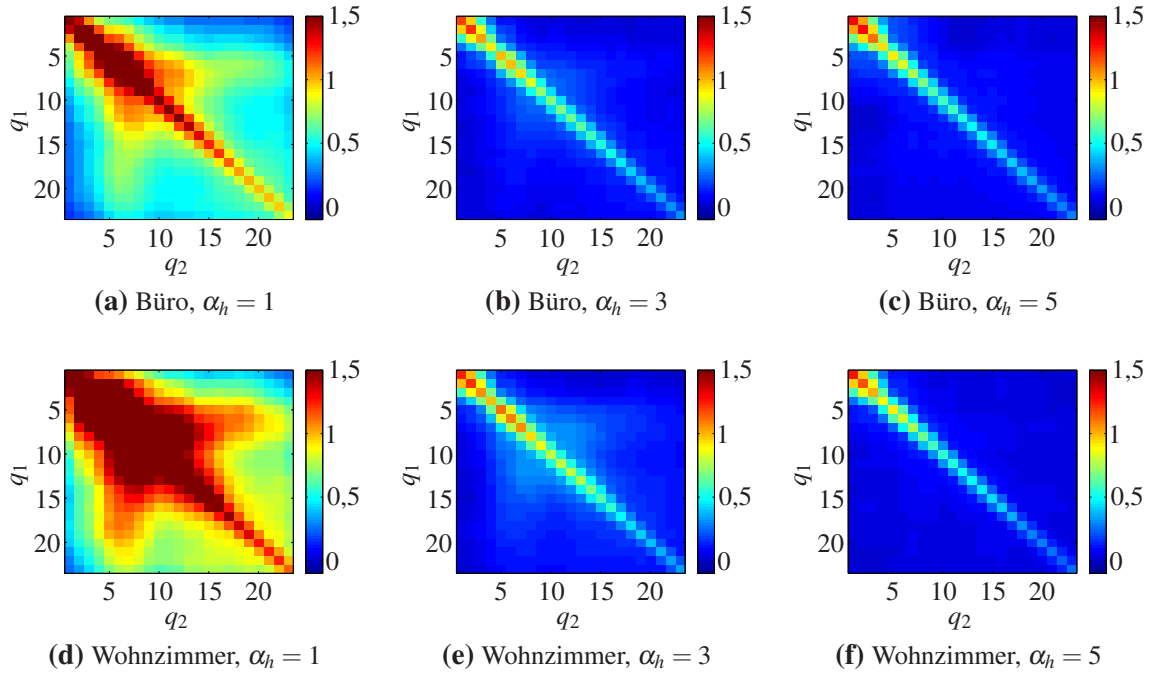


(a) Büro ($\hat{T}_{60} = 0{,}35\,\mathrm{s}$) **(b)** Wohnzimmer ($\hat{T}_{60} = 0{,}45\,\mathrm{s}$)

Abbildung 6.3.: *Approximative log-MEL-spekrale Repräsentationen der RIAs $\mu_{\check{h}_{m,q}}$ der beiden virtuellen Räume der AURORA5-Datenbank.*

Approximation erhält man durch einen Vergleich mit den entsprechenden wahren raumspezifischen log-MEL-spektralen Repräsentationen in Abb. A.2 im Anhang.

Weiterhin sind in Abb. 6.4 die Kovarianzmatrizen des Beobachtungsfehlers $\hat{\boldsymbol{\Sigma}}_{\hat{\mathbf{v}}^{(\mathrm{s})}}$ beispielhaft für die AURORA5-Datenbank für $\alpha_h = 1, 3, 5$ dargestellt, wobei q_1 den Zeilen- und q_2 den Spaltenindex kennzeichnet. Es lässt sich beobachten, dass mit zunehmenden Werten von



(a) Büro, $\alpha_h = 1$ **(b)** Büro, $\alpha_h = 3$ **(c)** Büro, $\alpha_h = 5$

(d) Wohnzimmer, $\alpha_h = 1$ **(e)** Wohnzimmer, $\alpha_h = 3$ **(f)** Wohnzimmer, $\alpha_h = 5$

Abbildung 6.4.: *Empirisch berechnete Kovarianzmatrizen des Beobachtungsfehlers $\hat{\boldsymbol{\Sigma}}_{\hat{\mathbf{v}}^{(\mathrm{s})}}$ ermittelt auf der AURORA5-Datenbank für die beiden untersuchten virtuellen Räume für verschiedene Werte von α_h.*

α_h nicht nur die Beträge der Werte der Diagonalelemente abnehmen, sondern insbesondere die der Nebendiagonalelemente. Dieses lässt sich darauf zurückführen, dass bedingt durch die Art der Berechnung der log-MEL-spektralen Merkmalsvektoren hauptsächlich Korrelationen zwischen benachbarten Vektorkomponenten auftreten, die durch die Überlappung benachbarter MEL-Bänder verursacht werden. Motiviert durch die approximativ diagonale Gestalt der Kovarianzmatrizen $\hat{\boldsymbol{\Sigma}}_{\hat{\mathbf{v}}^{(\mathrm{s})}}$ für größere Werte von α_h wurden für die weiteren Experimente stets diagonale Kovarianzmatrizen verwendet. Durch hier nicht weiter beschriebene Experimente wurde zudem festgestellt, dass die Verwendung von voll besetzten Kovarianzmatrizen im Vergleich zu diagonalen Kovarianzmatrizen zu einer insgesamt schlechteren Leistungsfähigkeit der Merkmalsverbesserung führte, auf die durch eine höhere Wortfehlerrate bei der anschließenden Erkennung geschlossen wurde.

Abbildung 6.5 zeigt die Histogramme für ausgewählte Komponenten $\hat{v}_{m,q}^{(\mathrm{s})}$ des Beobachtungsfehlervektors $\hat{\mathbf{v}}_m^{(\mathrm{s})}$, welche auf der modifizierten AURORA4-Datenbank für das Wohnzimmer ermittelt wurden, samt den entsprechenden Approximationen durch GAUSS-Verteilungsdichtefunktionen. Es lässt sich erkennen, dass trotz einer geringen Linksschiefe der Histogramme die vorgenommenen Näherungen durchaus sinnvoll sind. Da sich sowohl für das Büro als auch für die AURORA5-Datenbank ähnliche Verläufe ergaben, sind die Resul-

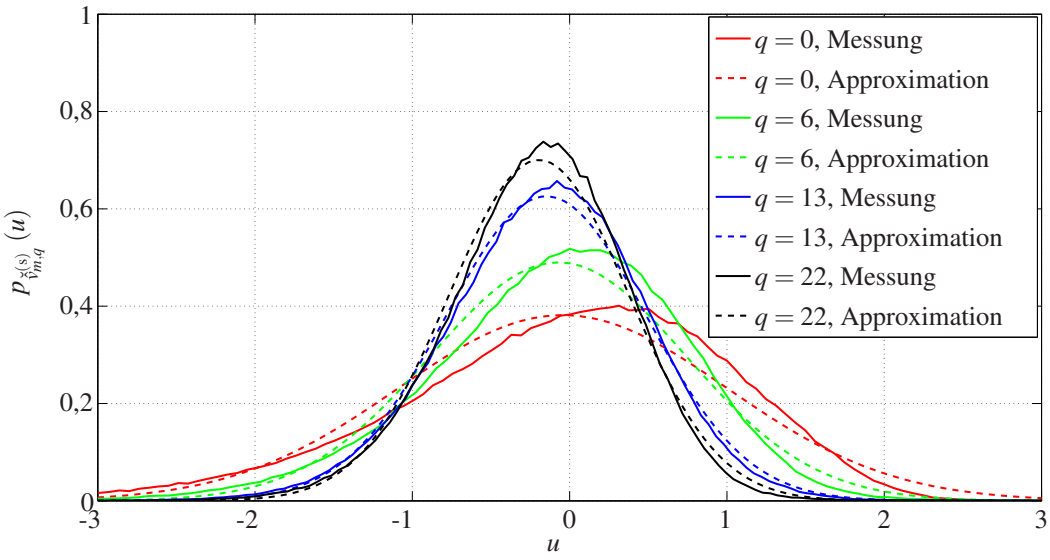
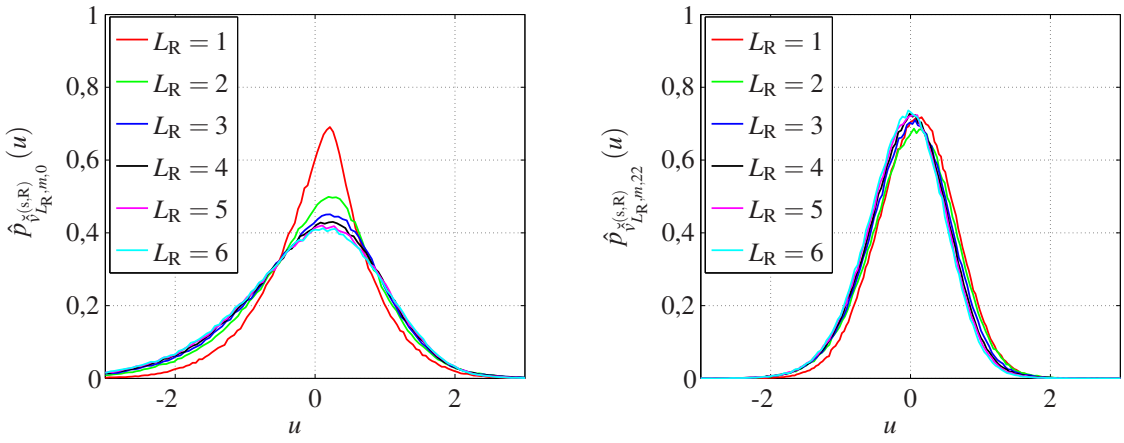


Abbildung 6.5.: *Empirisch berechnete normierte Histogramme ausgewählter Komponenten $\hat{v}^{(s)}_{m,q}$ des Beobachtungsfehlervektors für das Wohnzimmerszenario der modifizierten AURORA4-Datenbank sowie zugehörige Approximationen durch* GAUSS*-Verteilungsdichtefunktionen.*

tate hier nicht explizit aufgeführt.

In einem weiteren Experiment wurde der Beobachtungsfehler $\hat{\mathbf{v}}^{(s,R)}_{m,L_R}$ unter Verwendung des rekursiven Beobachtungsmodells und derselben Trainingsdaten wie im Experiment zuvor für die beiden untersuchten Räume und beide Datenbanken berechnet. Dabei ist zu beachten, dass für das rekursive Beobachtungsmodell keine Schätzung der Länge $\hat{L}_H$ der Repräsentation der RIA im log-MEL-spektralen Bereich erforderlich ist. Statt dessen muss eine Rekursionslänge L_R vorgegeben werden, welche den Beobachtungsfehler beeinflusst. In Abb. 6.6 sind exemplarisch die normierten Histogramme des Beobachtungsfehlers $\hat{v}^{(s,R)}_{L_R,m,q}$, welche für das Wohnzimmer auf der AURORA4-Datenbank bestimmt wurden und als Schätzungen der entsprechenden Verteilungsdichtefunktionen angesehen werden können, für unterschiedliche Rekursionslängen L_R und zwei ausgewählte MEL-Bänder ($q = 0$ und $q = 22$) illustriert. Wie in Abb. 6.6a am Beispiel für $q = 0$ ersichtlich wird, zeichnen sich die normierten Histogramme des Beobachtungsfehlers für niedrige MEL-Bänder q und sehr kleine Werte von L_R durch eine geringe Steilgipfligkeit und Linksschiefe aus. Das Ausmaß der Steilgipfligkeit wird jedoch mit der Erhöhung der Rekursionslänge L_R reduziert. Für höhere MEL-Bänder, d.h. für $q > 10$, treten beide Phänomene nur in einer sehr geringfügigen Ausprägung auf. Zudem besteht dann nur noch ein sehr kleiner Unterschied zwischen den Histogrammen für unterschiedliche Werte von L_R, wie am Beispiel für $q = 22$ in Abb. 6.6b deutlich wird.

Da die empirisch berechneten Kovarianzmatrizen des Beobachtungsfehlers $\hat{\mathbf{v}}^{(s,R)}_{L_R,m}$ bei der Verwendung der rekursiven Beobachtungsfunktion eine ähnliche Form wie jene im Fall der nichtrekursiven Beobachtungsfunktion aufwiesen, wurden diese ebenfalls für alle weiteren Experimente durch Diagonalmatrizen approximiert.



(a) $q = 0$ **(b)** $q = 22$

Abbildung 6.6.: *Empirisch berechnete normierte Histogramme des Beobachtungsfehlers $\hat{v}^{(s,R)}_{L_R,m,q}$ für unterschiedliche Rekursionslängen L_R und zwei ausgewählte* MEL*-Bänder ($q = 0$ und $q = 22$), ermittelt für das Wohnzimmer auf der modifizierten AURORA4-Datenbank.*

6.5. Ergebnisse zur Merkmalsenthallung

In diesem Abschnitt werden Ergebnisse des vorgestellten Verfahrens zur Merkmalsverbesserung für den Fall präsentiert, in dem keine Hintergrundstörungen im Mikrophonsignal präsent sind. Es wird daher zunächst nur die Leistungsfähigkeit des Verfahrens im Bezug auf die Enthallung von akustischen Merkmalen experimentell untersucht.

Dazu wurde in einem ersten Experiment die Leistungsfähigkeit verschiedener Modellkombinationsalgorithmen analysiert. Gleichzeitig wurden dabei die Anzahl I der A-priori-Teilmodelle sowie die Anzahl $L_C - 1$ der vorhergehenden sauberen Merkmalsvektoren innerhalb des Zustandsvektors variiert. Das A-priori-Modell wurde dabei mit Hilfe der in Kap. 5.1.4 beschriebenen iterativen Modellspaltung trainiert. Die Anzahl L_{EM} der *EM*-Iterationen nach jeder Modellspaltung wurde so gewählt, dass die mittlere relative Verbesserung der Likelihoodfunktion einen Wert von 10 zum ersten Mal unterschritt, d.h. $\delta_{\mathscr{L}}^{(L_{EM})} < 10$ und $\delta_{\mathscr{L}}^{(l)} \geq 10 \;\; \forall l < L_{EM}$.

Die resultierenden Wortfehlerraten für die AURORA5-Datenbank sind in Tab. 6.5 angegeben. Bei der Betrachtung der Ergebnisse fällt auf, dass durchgehend für alle Modellkombinationsalgorithmen bis auf *GPB1* für $I = 16$ und kleine Werte von L_C eine Verringerung der Wortfehlerrate erzielt worden ist. Insbesondere ist zu beobachten, dass sich die Resultate mit steigenden Werten von L_C zunächst deutlich verbessern, wobei die Verbesserung monoton abnimmt und bei etwa $L_C = 6$ eine Sättigung auftritt. Der Grund für diese Verbesserung liegt wie bereits weiter oben erwähnt darin, dass durch eine Vergrößerung von L_C eine stärkere Berücksichtigung der Zukunft stattfindet, die ihre Wirkung im Zusammenhang mit dem dispersiven Effekt des Nachhalls entfaltet. Das Auftreten einer Sättigung der Wortfehlerrate bestärkt diese Interpretation, da die zeitliche Ausdehnung der Verschmierung beschränkt ist.

Im Hinblick auf das A-priori-Modell ist bemerkenswert, dass bereits mit einem einzigen Teilmodell, d.h. $I = 1$, die Wortfehlerrate um bis zu etwa 70 % bei beiden Räumen reduziert

Tabelle 6.5.: *Wortfehlerraten λ_w [%] erzielt mit Hilfe der Merkmalsverbesserung auf der AURORA5-Datenbank.*

			Büro					Wohnzimmer				
			I					*I*				
			1	2	4	8	16	1	2	4	8	16
GPB1	L_C	1	4,47	4,75	4,35	5,01	18,70	11,28	12,62	11,35	12,70	25,54
		2	2,97	2,73	2,64	2,73	10,81	6,69	6,59	6,15	6,31	14,29
		3	2,53	2,41	2,30	2,27	7,41	5,62	5,34	4,80	4,86	9,76
		4	2,38	2,19	2,09	2,07	5,12	5,00	4,68	4,16	4,05	7,44
		5	2,17	2,07	1,93	1,97	3,96	4,58	4,26	3,72	3,67	6,23
		6	2,09	1,99	1,87	1,91	3,24	4,24	4,04	3,52	3,43	5,86
IMM	L_C	1	4,47	4,07	3,81	4,05	5,20	11,28	10,87	9,30	9,16	10,12
		2	2,97	2,57	2,69	2,71	3,31	6,69	6,16	6,07	5,94	6,36
		3	2,53	2,23	2,35	2,37	2,60	5,62	4,97	4,96	4,80	4,97
		4	2,38	2,12	2,17	2,20	2,28	5,00	4,32	4,28	4,04	4,08
		5	2,17	1,98	1,99	2,01	2,09	4,58	3,91	3,82	3,61	3,57
		6	2,09	1,93	1,97	1,93	1,93	4,24	3,77	3,61	3,40	3,32
GPB2	L_C	1	4,47	3,80	3,76	–	–	11,28	10,10	9,09	–	–
		2	2,97	2,55	2,58	–	–	6,69	6,19	6,23	–	–
		3	2,53	2,26	2,34	–	–	5,62	5,06	4,98	–	–
		4	2,38	2,12	2,17	–	–	5,00	4,36	4,30	–	–
		5	2,17	1,97	2,04	–	–	4,58	3,95	3,83	–	–
		6	2,09	1,92	1,96	–	–	4,24	3,85	3,64	–	–

werden kann. Dieses entspricht einer relativen Reduktion derjenigen Fehler, die durch den Nachhall verursacht worden sind, um etwa 75 %. In diesem Fall ist im Grunde keine Modellkombination erforderlich, so dass für die Merkmalsverbesserung ein gewöhnliches *IEKF* eingesetzt werden kann.

Die Vergrößerung der Anzahl der Teilmodelle I wirkt sich nicht immer positiv auf die Reduktion der Wortfehlerrate aus. Insbesondere ist die Tendenz zu beobachten, dass die Vergrößerung der Anzahl der Teilmodelle erst bei einem genügend groß gewählten Wert von L_C sinnvoll ist. Außerdem lässt sich feststellen, dass sie bei den im Vergleich zur *GPB1*-Schätzung komplizierteren und genaueren Modellkombinationsverfahren wie der *IMM*- und *GPB2*-Schätzung eher zu einer Reduktion der Wortfehlerrate führt. Aufgrund der Suboptimalität aller drei verwendeten Modellkombinationsalgorithmen kann eine Reduktion der Wortfehlerrate durch die Vergrößerung von I jedoch in keinem einzigen Fall gewährleistet werden. Einen weiteren Grund dafür, dass die Erhöhung der Teilmodellanzahl nicht immer mit einer Verringerung der Wortfehlerrate einhergeht, bildet die Tatsache, dass das zum Training des *SLDM* angewendete Kriterium der Maximierung der Loglikelihoodfunktion damit nicht unbedingt im Einklang steht. Es zeigt sich weiterhin, dass das *GPB2*-Verfahren trotz des deutlich höheren Aufwandes, der etwa quadratisch mit der Anzahl der Teilmodelle I wächst, keinen sichtbaren Vorteil gegenüber den *GPB1*- und *IMM*-Verfahren bietet.

Im Sinne einer vorsichtigen Beurteilung der Wortfehlerraten ist zu bemerken, dass in dieser Arbeit keine Signifikanztests bezüglich des Unterschieds von Wortfehlerraten unter-

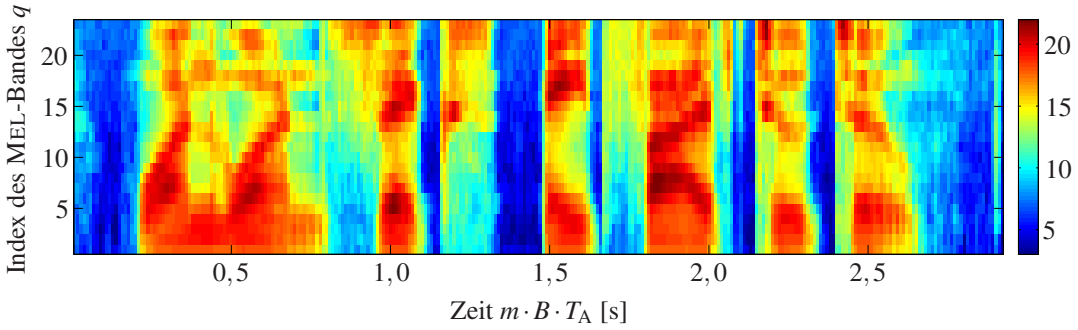
schiedlicher Verfahren durchgeführt wurden. Aus den Ergebnissen soll hier lediglich auf grobe Tenzenden geschlossen werden ohne dabei marginale, jedoch eventuell signifikante, Unterschiede zu interpretieren. In Kap. A.4 im Anhang wird aber dennoch zusätzlich darauf eingegangen, wie eine sehr grobe Beurteilung der Signifikanz der Unterschiede zweier Wortfehlerraten nur mit Hilfe der Erkennungsergebnisse berechnet werden kann und welche Aspekte eigentlich für eine genauere Betrachtung berücksichtigt werden müssen.

Um einen qualitativen Eindruck von der Leistungsfähigkeit der Merkmalsenthallung zu vermitteln, sind in Abb. 6.7 die Trajektorien der LMSK-Vektoren jeweils für ein beispielhaftes sauberes Sprachsignal und dessen verhallte Version (im Wohnzimmer) sowie die entsprechenden Trajektorien der verbesserten LMSK-Vektoren jeweils für $L_{\mathrm{C}} = 2$ und $L_{\mathrm{C}} = 6$ abgebildet. Es lässt sich deutlich erkennen, dass die mit Hilfe der Merkmalsverbesserung die Auswirkungen der durch den Nachhall bedingten zeitlichen Dispersion merkbar reduziert werden können. Beispielsweise ist der Glottalschlag bei der Aussprache der Ziffer "*six*" bei etwa 1,2 s, der in der Trajektorie der log-MEL-spektralen Merkmale $y_{m,q}^{(\mathrm{s})}$ des verhallten Sprachsignals in Abb. 6.7b vollkommen verdeckt ist, in den Trajektorien der verbesserten log-MEL-spektralen Merkmale $\hat{x}_{m,q}^{(\mathrm{s})}$ in Abb. 6.7c und Abb. 6.7d teilweise wieder erkennbar. Insbesondere wird durch Abb. 6.7c und Abb. 6.7d veranschaulicht, dass der Verlauf der Trajektorien verbesserten log-MEL-spektralen Merkmale $\hat{x}_{m,q}^{(\mathrm{s})}$ in der Regel mit wachsenden Werten von L_{C} zunehmend glatter wird, wobei die Auswirkungen der zeitlichen Dispersion weiter abnehmen.

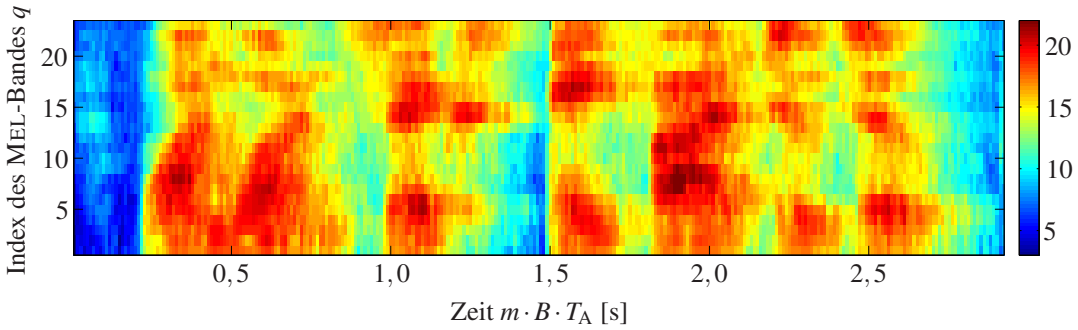
Zur Beurteilung des Rechenaufwandes des Verfahrens ist zu bemerken, dass die unterschiedlichen Inferenzalgorithmen (siehe Kap. 5.3) in C/C++ implementiert wurden und auf einem Rechner mit einem *Intel Core i7/*2,67 GHz-Prozessor ausgeführt wurden. Die Echtzeitfaktoren für unterschiedliche Parameterwahlen für das Wohnzimmer sind in Tab. 6.6 aufgeführt. Sie zeigen, dass das Verfahren auch aus Sicht des Rechenaufwandes echtzeitfähig ist, wobei wie bereits erwähnt eine Latenz der Dauer von $L_{\mathrm{C}} - 1$ Segmenten zu berücksichtigen ist. Insbesondere soll darauf hingewiesen werden, dass die Rechendauer für $L_{\mathrm{C}} > 1$ mit Hilfe einer parallelen Berechnung der teilmodellspezifischen Inferenzen deutlich reduziert werden kann.

Für die Erkennungsaufgabe mit einem großen Vokabular zeigte sich ein ähnliches Verhalten bezüglich der Wahl des Modellkombinationsalgorithmus sowie der Anzahl der Teilmodelle I und des Wertes von L_{C}. Die Wortfehlerraten für die modifizierte AURORA4-Datenbank für $I = 1$ in Tab. 6.7 veranschaulichen die Bedeutung der Erhöhung von L_{C} für die Leistungsfähigkeit des Verfahrens zur Merkmalsverbesserung. Die Wortfehlerrate konnte bei beiden Räumen bis um etwa 40 % reduziert werden. Dieses entspricht einer relativen Reduktion der durch den Nachhall verursachten Fehler um etwa 55 % beim Büro und um etwa 50 % beim Wohnzimmer. Die relativen Verbesserungen fielen erwartungsgemäß geringer aus als für die Ziffernkettenerkennung, da nach der Merkmalsverbesserung verbliebene Fehler aufgrund der hohen Komplexität der Erkennungsaufgabe schwerwiegendere Auswirkungen hatten.

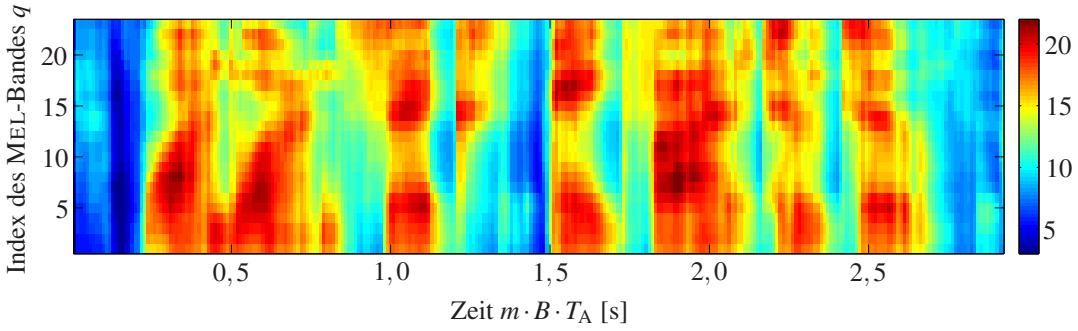
Die Wortfehlerraten für $I > 1$ und unterschiedliche Modellkombinationsalgorithmen und Werte von L_{C} lassen sich für das Büro aus Tab. 6.8 und für das Wohnzimmer aus Tab. 6.9 entnehmen. Hervorzuheben ist die Tatsache, dass alle drei Modellkombinationsalgorithmen auch in diesem Fall ähnliche Ergebnisse lieferten, obwohl der Aufwand für das *GPB2*-Verfahren deutlich größer war (siehe Tab. 6.6). Durch eine Vergrößerung der Anzahl der



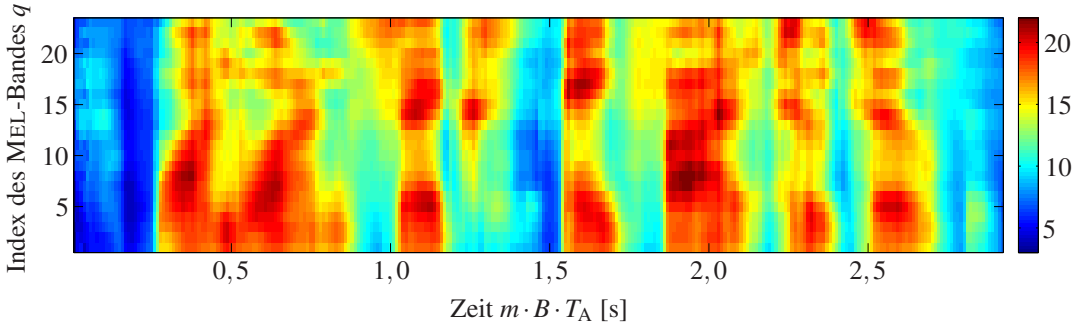
(a) Trajektorie der log-MEL-spektralen Merkmale $x_{m,q}^{(s)}$ des sauberen Sprachsignals



(b) Trajektorie der log-MEL-spektralen Merkmale $y_{m,q}^{(s)}$ des verhallten Sprachsignals



(c) Trajektorie der verbesserten log-MEL-spektralen Merkmale $\hat{x}_{m,q}^{(s)}$ für $L_C = 2$



(d) Trajektorie der verbesserten log-MEL-spektralen Merkmale $\hat{x}_{m,q}^{(s)}$ für $L_C = 6$

Abbildung 6.7.: *Trajektorien der log-MEL-spektralen Merkmale eines beispielhaften Sprachsignals der AURORA5-Datenbank zugehörig zu der Ziffernkettenäußerung "one, one, six, eight, five, two, two".*

Tabelle 6.6.: *Echtzeitfaktoren für die Merkmalsverbesserung für das Wohnzimmer ($\hat{L}_H = 25$).*

			I				
			1	2	4	8	16
GPB1	L_C	1	0,03	0,06	0,12	0,23	0,46
		2	0,04	0,08	0,17	0,32	0,63
		3	0,06	0,11	0,22	0,44	0,88
		4	0,08	0,15	0,30	0,59	1,17
		5	0,10	0,19	0,38	0,76	1,51
		6	0,12	0,24	0,48	0,95	1,92
IMM	L_C	1	0,03	0,06	0,12	0,24	0,49
		2	0,04	0,08	0,17	0,34	0,72
		3	0,06	0,12	0,24	0,49	1,07
		4	0,08	0,16	0,32	0,67	1,50
		5	0,10	0,20	0,42	0,88	2,02
		6	0,12	0,26	0,54	1,15	2,66
GPB2	L_C	1	0,03	0,12	0,47	–	–
		2	0,04	0,16	0,66	–	–
		3	0,06	0,23	0,91	–	–
		4	0,08	0,31	1,21	–	–
		5	0,10	0,39	1,57	–	–
		6	0,13	0,50	1,94	–	–

Tabelle 6.7.: *Fehlerraten [%] erzielt mit Hilfe der Merkmalsverbesserung auf der modifizierten AURORA4-Datenbank für $I = 1$.*

		Raum							
		Büro				Wohnzimmer			
		λ_{Subst}	λ_{Ausl}	λ_{Einf}	λ_w	λ_{Subst}	λ_{Ausl}	λ_{Einf}	λ_w
L_C	1	33,85	3,57	7,22	**44,64**	54,44	7,00	8,62	**70,06**
	2	26,74	2,76	7,85	**37,35**	42,50	3,76	9,72	**55,99**
	3	23,83	2,65	7,00	**33,48**	38,45	3,68	9,76	**51,90**
	4	22,14	2,69	6,56	**31,38**	36,61	3,35	8,58	**48,55**
	5	21,44	2,87	5,97	**30,28**	33,81	3,79	8,62	**46,22**
	6	20,77	2,58	5,56	**28,91**	32,38	4,16	8,21	**44,75**

Tabelle 6.8.: *Fehlerraten [%] erzielt mit Hilfe der Merkmalsverbesserung auf der modifizierten AURORA4-Datenbank für das Büro.*

			I = 2				I = 4			
			λ_{Subst}	λ_{Ausl}	λ_{Einf}	λ_w	λ_{Subst}	λ_{Ausl}	λ_{Einf}	λ_w
GPB1	L_{C}	1	35,80	3,87	8,21	**47,88**	33,55	3,46	8,43	**45,45**
		2	25,86	2,54	7,44	**35,84**	25,01	2,50	7,55	**35,06**
		3	22,80	2,58	7,07	**32,45**	22,50	2,58	7,15	**32,23**
		4	21,40	2,84	6,22	**30,46**	20,99	2,32	6,37	**29,69**
		5	20,96	2,80	5,82	**29,58**	20,66	2,28	5,78	**28,73**
		6	19,85	2,91	5,45	**28,21**	20,11	2,62	5,41	**28,14**
IMM	L_{C}	1	33,48	4,20	7,92	**45,60**	30,31	3,28	8,03	**41,62**
		2	26,08	2,69	7,81	**36,57**	25,38	2,28	7,73	**35,40**
		3	23,39	2,28	6,74	**32,41**	23,02	2,43	7,18	**32,63**
		4	21,80	2,69	6,26	**30,76**	21,44	2,65	6,52	**30,61**
		5	21,07	2,80	5,52	**29,39**	20,96	2,36	5,41	**28,73**
		6	20,52	2,69	5,23	**28,43**	20,07	2,54	5,16	**27,77**
GPB2	L_{C}	1	32,41	3,68	7,88	**43,98**	31,90	2,95	7,51	**42,36**
		2	26,52	2,62	7,66	**36,80**	25,49	2,39	7,44	**35,32**
		3	23,31	2,47	6,96	**32,74**	22,91	2,28	7,11	**32,30**
		4	21,84	2,69	6,48	**31,01**	21,73	2,32	6,11	**30,17**
		5	20,92	2,69	5,56	**29,17**	21,33	2,36	5,64	**29,32**
		6	20,26	2,58	5,34	**28,18**	20,44	2,69	5,08	**28,21**

(a) $I = 2,4$

			I = 8				I = 16			
			λ_{Subst}	λ_{Ausl}	λ_{Einf}	λ_w	λ_{Subst}	λ_{Ausl}	λ_{Einf}	λ_w
GPB1	L_{C}	1	31,93	3,24	8,77	**43,94**	31,16	2,58	8,29	**42,03**
		2	25,49	2,32	7,62	**35,43**	25,27	2,25	7,85	**35,36**
		3	22,50	2,25	6,63	**31,38**	22,73	2,25	6,56	**31,53**
		4	20,81	2,47	6,11	**29,39**	21,18	2,32	5,93	**29,43**
		5	20,85	2,32	5,64	**28,80**	20,41	2,06	5,30	**27,77**
		6	20,77	2,10	5,41	**28,29**	20,41	2,06	5,12	**27,59**
IMM	L_{C}	1	29,32	3,31	8,58	**41,22**	28,77	2,62	8,84	**40,22**
		2	25,75	2,32	8,29	**36,35**	25,78	2,17	8,10	**36,06**
		3	23,13	2,50	7,29	**32,93**	23,50	2,54	7,85	**33,89**
		4	21,58	2,54	7,22	**31,34**	22,65	2,50	6,78	**31,93**
		5	20,74	2,32	5,82	**28,88**	21,47	2,03	5,97	**29,47**
		6	20,66	2,54	5,82	**29,02**	21,44	2,06	5,71	**29,21**

(b) $I = 8,16$

Tabelle 6.9.: *Fehlerraten [%] erzielt mit Hilfe der Merkmalsverbesserung auf der modifizierten AURORA4-Datenbank für das Wohnzimmer.*

			I = 2				I = 4			
			λ_{Subst}	λ_{Ausl}	λ_{Einf}	λ_w	λ_{Subst}	λ_{Ausl}	λ_{Einf}	λ_w
GPB1	L_{C}	1	55,51	8,36	8,40	**72,27**	54,40	6,78	9,32	**70,50**
		2	41,69	4,24	9,65	**55,58**	41,80	3,65	10,57	**56,02**
		3	37,50	3,72	9,76	**50,98**	36,80	3,54	8,99	**49,32**
		4	34,99	3,24	8,47	**46,70**	33,30	3,39	8,21	**44,90**
		5	32,60	3,98	8,55	**45,12**	31,31	3,39	8,18	**42,87**
		6	30,90	4,01	7,92	**42,84**	29,06	3,65	7,51	**40,22**
IMM	L_{C}	1	55,99	6,30	9,02	**71,31**	49,94	4,79	9,98	**64,71**
		2	41,29	3,65	10,09	**55,03**	39,26	3,50	10,31	**53,08**
		3	37,35	3,20	9,24	**49,80**	36,39	3,28	9,43	**49,10**
		4	33,26	3,13	8,51	**44,90**	32,71	3,35	9,47	**45,52**
		5	31,71	3,68	8,58	**43,98**	30,94	3,20	8,47	**42,62**
		6	30,39	3,90	7,73	**42,03**	29,10	3,61	7,73	**40,44**
GPB2	L_{C}	1	54,03	5,56	10,28	**69,87**	50,50	5,30	10,98	**66,78**
		2	42,62	3,54	9,80	**55,95**	40,15	3,68	11,09	**54,92**
		3	37,72	3,13	9,24	**50,09**	36,43	3,43	9,36	**49,21**
		4	32,89	3,24	8,43	**44,57**	32,60	3,35	8,95	**44,90**
		5	31,86	3,54	8,36	**43,76**	31,45	3,09	8,91	**43,46**
		6	31,12	3,61	8,14	**42,87**	29,36	3,65	7,85	**40,85**

(a) $I = 2, 4$

			I = 8				I = 16			
			λ_{Subst}	λ_{Ausl}	λ_{Einf}	λ_w	λ_{Subst}	λ_{Ausl}	λ_{Einf}	λ_w
GPB1	L_{C}	1	53,08	5,19	10,57	**68,84**	50,94	5,23	10,28	**66,45**
		2	39,67	3,87	9,94	**53,48**	39,85	3,39	10,20	**53,44**
		3	35,21	3,35	8,58	**47,15**	35,14	3,54	9,10	**47,77**
		4	31,90	3,46	8,14	**43,50**	31,49	3,50	8,14	**43,13**
		5	30,17	3,24	7,51	**40,92**	29,76	3,43	7,37	**40,55**
		6	28,66	3,39	7,26	**39,30**	28,84	3,54	6,96	**39,34**
IMM	L_{C}	1	49,69	4,71	10,17	**64,57**	47,77	4,24	11,27	**63,28**
		2	39,37	3,68	10,13	**53,19**	38,78	3,35	10,64	**52,78**
		3	35,95	3,20	9,02	**48,18**	34,81	2,98	9,69	**47,48**
		4	32,89	3,09	9,10	**45,08**	32,71	3,17	9,24	**45,12**
		5	30,64	3,13	8,43	**42,21**	30,64	3,24	8,73	**42,62**
		6	29,76	3,20	8,25	**41,22**	28,99	3,17	8,03	**40,18**

(b) $I = 8, 16$

Teilmodelle I konnte die Wortfehlerrate in den meisten Fällen geringfügig reduziert werden. Die erzielte Verbesserung bei der Nutzung vieler Teilmodelle steht dabei in einem sehr ungünstigen Verhältnis zum aufgebrachten Rechenaufwand.

Im Sinne eines vernünftigen Kompromisses zwischen Rechenaufwand und Leistungsfähigkeit wurden für die weiteren Experimente $I = 4$ Teilmodelle verwendet, wobei als Modellkombinationsalgorithmus die *IMM*-Methode diente. Da die Erhöhung von L_C in hohem Maße zur Reduktion der Wortfehlerrate beitrug und der Rechenaufwand im Verhältnis zur Verbesserung der Leistungsfähigkeit der Merkmalsverbesserung vertretbar anstieg, wurde für die folgenden Untersuchungen $L_\mathrm{C} = 6$ gewählt.

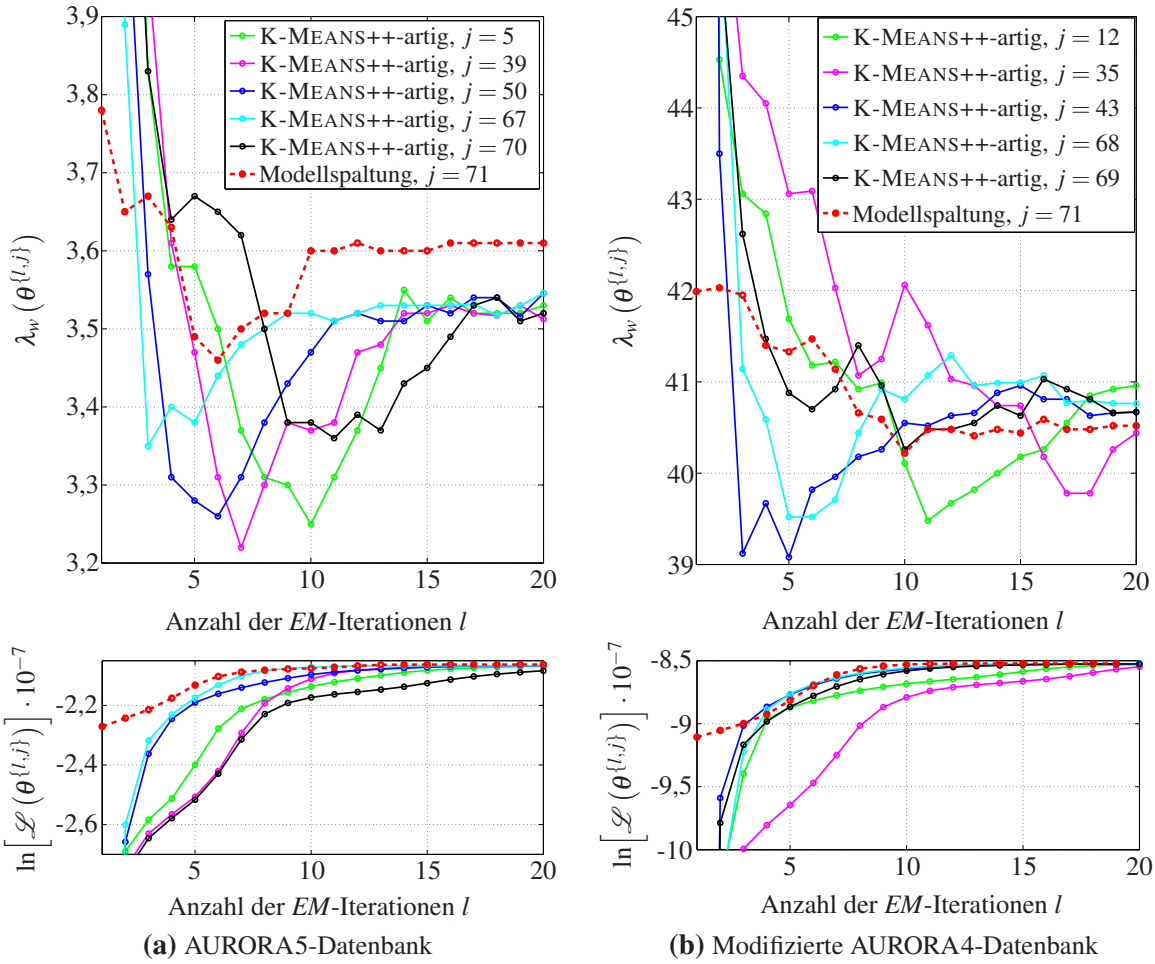
6.5.1. Einfluss des A-priori-Sprachmodells

In diesem Abschnitt wird der Einfluss des A-priori-Sprachmodells auf die Leistungsfähigkeit der Merkmalsverbesserung untersucht. Das Ziel liegt hierbei nicht in der Bestimmung eines optimalen Sprachmodells für die betrachteten Sprachdatenbanken, sondern vielmehr darin, den Einfluss der in Kap. 5.1.4 diskutierten Initialisierung und einzelner *EM*-Iterationen des Trainings des Sprachmodells sowie den Einfluss der Ordnung des Sprachmodells zu veranschaulichen.

Einfluss der Initialisierung

Zunächst wurden für beide Sprachdatenbanken für das Wohnzimmer jeweils 70 unterschiedliche initiale Parametermengen $\theta^{\{0,j\}}$ des A-priori-Sprachmodells mit $I = 4$ Teilmodellen mit der in Kap. 5.1.4 vorgeschlagenen K-MEANS++-artigen Methode bestimmt, wobei $j \in \{1,\ldots,70\}$ den Index der Initialisierung bezeichnet. Die Länge der aus den Trainingsdaten zufällig ausgewählten Segmente zur Bestimmung lokaler Teilmodelle wurde zu $L_\mathrm{S} = 10$ gewählt um einen vernünftigen Kompromiss zwischen Lokalität und Informationsgehalt der Teilmodelle zu gewährleisten. Die Konstante $\varepsilon_{P,\mathrm{REL}}$ zur Steuerung des Verwerfens unterrepräsentierter Teilmodelle wurde zu $0,01$ gesetzt (siehe Alg. 2). Anschließend wurde jedes dieser derart initialisierten Sprachmodelle unter Anwendung von jeweils 20 Iterationen des *EM*-Algorithmus trainiert. Dabei wurden die nach der l-ten Iteration berechneten Parametermengen $\theta^{\{l,j\}}$ jeweils zwischengespeichert, so dass eine Menge von insgesamt 1400 A-priori-Sprachmodellen resultierte. Außerdem wurde, wie bereits im letzten Abschnitt beschrieben, ein initiales A-priori-Sprachmodell $\theta^{\{0,71\}}$ mit $I = 4$ Teilmodellen mit Hilfe der Modellspaltung erzeugt und ebenfalls mit 20 Iterationen des *EM*-Algorithmus verfeinert. Mit jedem der beschriebenen A-priori-Sprachmodelle wurde eine Merkmalsverbesserung mit anschließender Spracherkennung durchgeführt.

In Abb. 6.8 sind für ausgewählte Indizes j der initialen Modellparameter die Verläufe der Wortfehlerrate $\lambda_w\left(\theta^{\{l,j\}}\right)$ sowie mit 10^{-7} skalierte Werte der Loglikelihoodfunktion $\ln\left[\mathscr{L}\left(\theta^{\{l,j\}}\right)\right]$ in Abhängigkeit von der Anzahl der *EM*-Iterationen l für die AURORA5- und die modifizierte AURORA4-Datenbank dargestellt. Bei allen mit der K-MEANS++-artigen Methode initialisierten Parametermengen ließ sich feststellen, dass die Wortfehlerrate innerhalb der ersten 3 *EM*-Iterationen beträchtlich abnahm. Dieses ist darauf zurückzuführen, dass die Initialisierung lediglich mit Hilfe weniger zufällig ausgewählter Merkmalsvektorsequenzen erfolgte, so dass die entsprechenden A-priori-Modelle lediglich eine sehr



(a) AURORA5-Datenbank **(b)** Modifizierte AURORA4-Datenbank

Abbildung 6.8.: *Wortfehlerraten $\lambda_w\left(\theta^{\{l,j\}}\right)$ sowie mit 10^{-7} skalierte Werte der Loglikelihoodfunktion $\ln\left[\mathscr{L}\left(\theta^{\{l,j\}}\right)\right]$ in Abhängigkeit von der Anzahl l der für das Training des A-priori-Sprachmodells verwendeten EM-Iterationen für beispielhaft ausgewählte initiale Parametermengen $\theta^{\{0,j\}}$ für das Wohnzimmer.*

lokale anstatt einer globalen Charakterisierung der Sprache boten. Innerhalb der ersten *EM*-Iterationen erfolgte eine Anpassung der lokalen Modelle an die globalen Daten, was sich durch einen enormen Anstieg der Werte der Loglikelihoodfunktion bemerkbar machte. Daraus lässt sich ableiten, dass eine gute Modellierung der Sprachdaten durch das A-priori-Sprachmodell in gewisser Weise zu einer niedrigen Wortfehlerrate beiträgt.

Dass beide Kriterien jedoch nicht äquivalent sind, zeigt folgende Beobachtung, die in den meisten Experimenten gemacht wurde und mit Hilfe der Verläufe der Wortfehlerrate in Abb. 6.8a und Abb. 6.8b veranschaulicht werden soll. Nachdem die Wortfehlerrate nach einigen *EM*-Iterationen ihr Minimum erreichte, stieg sie danach in geringem Maße wieder an. Aus Sicht der Merkmalsverbesserung wird eine Beschreibung der Sprachdaten mit Hilfe von Teilmodellen $\mathfrak{S}_i$ mit geringen Unsicherheiten, welche sich in kleinen Spektralradien der Kovarianzmatrizen $\mathbf{V}_i$ ausdrücken, favorisiert. Eine mögliche Ursache für den erneuten Anstieg der Wortfehlerrate könnte darin bestehen, dass diese Nebenbedingung bei der Durchführung

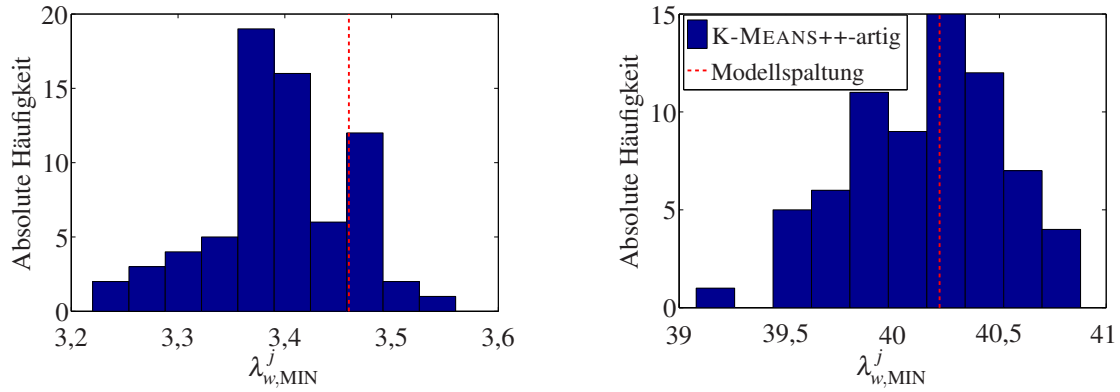
des *EM*-Algorithmus zur Maximierung der Likelihoodfunktion nicht beachtet wurde. Ein weiterer Grund liegt sicherlich in der Suboptimalität der Modellkombination, deren negativer Einfluss auf die Wortfehlerrate auch vom A-priori-Sprachmodell abhängt.

Ein ziemlich unerwünschter Effekt war dabei die Tatsache, dass die Anzahl der *EM*-Iterationen, nach denen das Minimum der Wortfehlerrate erreicht wurdw, stets unterschiedlich war. Diese Beobachtung weist auf das grundsätzliche Problem hin, dass sich während des Trainings des A-priori-Modells nicht vorhersagen lässt, welche Anzahl von *EM*-Iterationen eine minimale Wortfehlerrate im Nachhinein erzielen wird. Eine sinnvolle Lösung dieses Problems sollte bereits beim Optimierungskriterium für das Training ansetzen. Bedauerlicherweise ist die Formulierung eines geeigneten Kriteriums recht kompliziert und das Problem immer noch offener Forschungsgegenstand.

Nichtsdestotrotz zeigen die Histogramme der minimalen Wortfehlerrate

$$\lambda_{w,\mathrm{MIN}}^{j} := \min_{1 \le l \le 20} \lambda_w \left(\theta^{\{l,j\}} \right) \tag{6.5}$$

in Abb. 6.9, dass die Leistungsfähigkeit der Merkmalsverbesserung im Hinblick auf die Spracherkennung nur durch eine andere Art der Initialisierung des A-priori-Sprachmodells verbessert werden konnte, wenn auch nur geringfügig. Es ist außerdem davon auszugehen, dass die Initialisierung bei Sprachmodellen mit einer größeren Anzahl von Teilmodellen einen wesentlich größeren Einfluss besitzen wird.



(a) AURORA5-Datenbank **(b)** Modifizierte AURORA4-Datenbank

Abbildung 6.9.: *Histogramme der minimalen Wortfehlerrate $\lambda_{w,\mathrm{MIN}}^{j}$ für das Wohnzimmer.*

Insgesamt lässt sich schlussfolgern, dass die vorgeschlagene Art der Initialisierung bei Weitem noch keine optimale Lösung bietet, sondern lediglich Potentiale aufzeigt.

Einfluss der Modellordnung

Dieser Abschnitt widmet sich der Fragestellung, ob und inwieweit die Erhöhung der Ordnung L_{AR} des A-priori-Sprachmodells zu einer weiteren Reduktion der Wortfehlerrate beitragen kann. Durch die Erhöhung der Ordnung lassen sich Korrelationen zwischen zeitlich weiter auseinander liegenden Merkmalsvektoren der Sprache berücksichtigen, so dass der Prädiktionsfehler durch das A-priori-Sprachmodell prinzipiell verringert werden kann.

In einem ersten Experiment wurden A-priori-Sprachmodelle unterschiedlicher Ordnung L_{AR} bestehend aus jeweils nur einem Teilmodell, d.h. $I = 1$, für beide untersuchten Sprachdatenbanken berechnet und anschließend für die Merkmalsenthallung verwendet. Die sich ergebenden Wortfehlerraten bei der darauf folgenden Spracherkennung in Tab. 6.10 verdeutlichen, dass in den meisten Fällen eine deutliche Verbesserung der Leistungsfähigkeit hauptsächlich durch die Erhöhung der Ordnung von 1 auf 2 erzielt werden konnte. Eine

Tabelle 6.10.: *Fehlerraten [%] für verschiedene Ordnungen L_{AR} des A-priori-Sprachmodells bestehend aus einem Teilmodell, d.h. $I = 1$.*

		Raum Büro	Raum Wohnzimmer
		λ_w	
L_{AR}	1	2,09	4,24
	2	1,82	3,84
	3	1,73	3,75
	4	1,74	3,72

(a) AURORA5-Datenbank

		Raum Büro				Raum Wohnzimmer			
		λ_{Subst}	λ_{Ausl}	λ_{Einf}	λ_w	λ_{Subst}	λ_{Ausl}	λ_{Einf}	λ_w
L_{AR}	1	20,77	2,58	5,56	**28,91**	32,38	4,16	8,21	**44,75**
	2	21,80	2,62	6,30	**30,72**	31,38	3,83	8,36	**43,57**
	3	21,47	2,58	6,22	**30,28**	31,23	3,72	8,47	**43,43**
	4	21,14	2,54	6,22	**29,91**	31,12	3,76	8,62	**43,50**

(b) Modifizierte AURORA4-Datenbank

weitere Erhöhung der Ordnung hatte nur marginale Effekte, da sich die Güte der Prädiktion dadurch nur in geringem Maße verbessert. Denn die Korrelation zwischen zeitlich benachbarten Merkmalsvektoren der Sprache nimmt erheblich mit der Erhöhung des zeitlichen Abstandes ab.

In einem weiteren Experiment wurden *SLDM*s der Ordnung 2 und 3 bestehend aus $I = 4$ Modellen trainiert, wobei für das *EM*-Training dieselben initialen Parametermengen $\mathfrak{S}_i$ wie für das Training von *SLDM*s der Ordnung 1 zugrunde gelegt wurden. Jedes A-priori-Sprachmodell wurde nach jeder einzelnen *EM*-Iteration zwischengespeichert und zur Merkmalsenthallung eingesetzt. Bei einem Vergleich der bei der anschließenden Spracherkennung resultierenden Wortfehlerraten mit denen, die mit *SLDM*s erster Ordnung erzielt worden sind, ließ sich kein Gewinn in der Leistungsfähigkeit durch die Erhöhung der Ordnung L_{AR} feststellen. Die Wortfehlerraten lagen im Gegenteil sogar geringfügig höher, obwohl die Werte der Loglikelihoodfunktion deutlich größer als im Falle von *SLDM*s der Ordnung 1 waren. Dieses Resultat deutet erneut auf die Suboptimalität des Kriteriums zum Training des *SLDM* im Hinblick für die Verwendung zur Merkmalsverbesserung als auch auf die Subopti-

malität der Modellkombinationsalgorithmen, deren Einsatz zur vernünftigen Beschränkung des Rechenaufwandes beim Vorhandensein mehrerer Teilmodelle notwendig ist, hin. Auf eine detaillierte Darstellung der experimentellen Ergebnisse für $L_{\mathrm{AR}} > 1$ und $I > 1$ wird hier verzichtet, da aus Sicht des Autors keine sinnvollen Erkenntnisse daraus gezogen werden können.

6.5.2. Einfluss des Beobachtungsmodells

In diesem Abschnitt soll der Einfluss des Beobachtungsmodells auf die Leistungsfähigkeit der Merkmalsenthallung untersucht werden. Dabei stehen zwei Aspekte im Vordergrund. Erstens soll experimentell bestätigt werden, dass mit der rekursiven Beobachtungsfunktion, die ursprünglich zur Reduktion des Rechen- und Speicheraufwands eingeführt wurde, ähnliche Resultate wie mit der nichtrekursiven Beobachtungsfunktion erzielt werden können. Zweitens soll die Robustheit des Verfahrens zur Merkmalsenthallung gegenüber Fehlschätzungen der Parameter des RIA-Modells analysiert werden.

Ergebnisse mit der rekursiven Beobachtungsfunktion

Die Merkmalsenthallung wurde nun unter Verwendung des rekursiven Beobachtungsmodells und der *IMM*-Schätzung für verschiedene Rekursionslängen L_{R} und Anzahl von Teilmodellen I durchgeführt. Dabei wurde die Anzahl L_{C} von aufeinander folgenden Merkmalsvektoren der sauberen Sprache innerhalb des Zustandsvektors stets gleich L_{R} gewählt.

Die resultierenden Wortfehlerraten sind für die AURORA5-Datenbank in Tab. 6.11 und für die modifizierte AURORA4-Datenbank in Tab. 6.12 aufgeführt. Es zeigt sich bei bei-

Tabelle 6.11.: *Wortfehlerraten λ_w [%] erzielt mit dem rekursiven Beobachtungsmodell und der IMM-Schätzung auf der AURORA5-Datenbank.*

		Raum	
		Büro	Wohnzimmer
L_{R}	1	2,83	8,12
	2	2,44	6,21
	3	2,53	5,79
	4	2,48	5,52
	5	2,48	5,30
	6	2,36	4,97
	7	2,35	4,76
	8	2,32	4,55

(a) $I = 1$

		Raum	
		Büro	Wohnzimmer
L_{R}	1	2,87	7,83
	2	2,17	5,24
	3	2,30	4,87
	4	2,30	4,72
	5	2,19	4,36
	6	2,17	4,13
	7	2,13	3,91
	8	2,17	3,76

(b) $I = 4$

den Datenbanken, dass die Wortfehlerrate deutlich mit der Erhöhung der Rekursionslänge bis zum Eintreten einer Sättigung bei etwa $L_{\mathrm{R}} = 8$ abnahm, was sich als Ergebnis der Verwendung von immer mehr Wissen aus der Zukunft zur Merkmalsenthallung erklären lässt. Insbesondere ist dieses Verhalten nicht auf einen geringer werdenden Beobachtungsfehler

Tabelle 6.12.: *Fehlerraten [%] erzielt mit dem rekursiven Beobachtungsmodell und der IMM-Schätzung auf der AURORA4-Datenbank.*

		Raum							
		Büro				Wohnzimmer			
		λ_{Subst}	λ_{Ausl}	λ_{Einf}	λ_w	λ_{Subst}	λ_{Ausl}	λ_{Einf}	λ_w
L_{R}	1	26,48	4,38	6,08	**36,94**	53,37	9,91	6,08	**69,36**
	2	22,06	3,02	5,67	**30,76**	41,40	6,41	6,63	**54,44**
	3	21,47	2,95	6,15	**30,57**	36,72	4,38	6,78	**47,88**
	4	21,73	2,62	6,37	**30,72**	35,17	3,76	7,44	**46,37**
	5	21,25	2,39	5,97	**29,61**	32,82	3,90	7,51	**44,24**
	6	21,07	2,87	5,67	**29,61**	32,89	4,01	7,81	**44,71**
	7	20,70	2,87	5,30	**28,88**	31,27	4,24	7,37	**42,87**
	8	20,11	2,91	5,08	**28,10**	31,05	4,20	6,89	**42,14**

(a) $I = 1$

		Raum							
		Büro				Wohnzimmer			
		λ_{Subst}	λ_{Ausl}	λ_{Einf}	λ_w	λ_{Subst}	λ_{Ausl}	λ_{Einf}	λ_w
L_{R}	1	25,64	3,79	5,82	**35,25**	54,84	7,88	7,51	**70,24**
	2	21,58	2,58	5,86	**30,02**	41,07	5,16	7,40	**53,63**
	3	20,99	2,50	6,26	**29,76**	34,36	3,98	7,62	**45,97**
	4	20,74	2,39	6,30	**29,43**	31,79	3,79	7,73	**43,31**
	5	19,93	2,50	5,75	**28,18**	31,49	3,54	7,59	**42,62**
	6	19,08	2,54	5,01	**26,63**	30,13	3,65	7,81	**41,58**
	7	19,08	2,73	5,41	**27,22**	29,80	3,43	7,66	**40,88**
	8	18,31	2,73	5,19	**26,22**	28,66	4,16	7,37	**40,18**

(b) $I = 4$

für wachsende Werte von L_R zurückzuführen, da sich die statistischen Eigenschaften des Beobachtungsfehlers für $L_R > 3$ nur noch unwesentlich ändern (siehe Abb. 6.6).

Bei einem Vergleich der Resultate mit denen für das nichtrekursive Beobachtungsmodell in Tab. 6.5, Tab. 6.7, Tab. 6.8 und Tab. 6.9 lässt sich feststellen, dass für sehr kleine Werte von $L_C = L_R$ das rekursive Beobachtungsmodell zu geringfügig besseren Ergebnisse führte. Für $L_R > 3$ lieferten jedoch beide Beobachtungsmodelle ähnliche Resultate.

Die Echtzeitfaktoren, die bei der Merkmalsenthallung mit dem rekursiven Beobachtungsmodell gemessen wurden, sind in Tab. 6.13 aufgelistet. Es zeigt sich, dass diese im Vergleich

Tabelle 6.13.: *Echtzeitfaktoren für die Merkmalsenthallung unter Verwendung des rekursiven Beobachtungsmodells.*

		L_R							
		1	2	3	4	5	6	7	8
I	1	0,02	0,03	0,05	0,07	0,09	0,12	0,15	0,19
	4	0,08	0,13	0,20	0,29	0,40	0,50	0,69	0,84

zur Merkmalsenthallung mit dem nichtrekursiven Beobachtungsmodell etwa um 0,01 für $I = 1$ und 0,04 für $I = 4$ absolut geringer sind. Der Gewinn bezüglich des Rechenaufwands ist also wie erwartet linear in der Anzahl I der Teilmodelle. Jedoch ist er im Vergleich zum Gesamtaufwand für größere Werte von L_R relativ gering.

Durch experimentelle Untersuchungen, deren Ergebnisse hier nicht explizit aufgeführt sind, konnte weiterhin festgestellt werden, dass sich die Erhöhung der Ordnung L_{AR} des *SLDM* ähnlich wie im Falle der nichtrekursiven Beobachtungsfunktion auswirkte. Insbesondere konnte bei der Verwendung eines einzigen linearen dynamischen Modells zur Modellierung der Sprache, d. h. $I = 1$, durch die Erhöhung der Ordnung von 1 auf 2 eine relativ große Reduktion der Wortfehlerrate erzielt werden. Bei der Verwendung mehrerer Teilmodelle, d. h. $I > 1$, führte die Erhöhung der Ordnung L_{AR} des *SLDM* hingegen sogar zu einer leichten Erhöhung der Wortfehlerrate. Der Grund dafür liegt hier, wie auch beim nicht rekursiven Modell, in der Suboptimalität des verwendeten Kriteriums zum Training des *SLDM* und in der Suboptimalität der Modellkombinationsalgorithmen.

Sensitivität gegenüber Fehlschätzungen der Modellparameter

In dieser Arbeit wird davon ausgegangen wird, dass die Schätzung der zur Merkmalsenthallung benötigten RIA-Parameter $\hat{T}_{60}$ und $\hat{\sigma}_h^2$ mit Hilfe von externen Verfahren geschieht. Da sie jedoch in der Regel fehlerbehaftet ist, wurde in einem weiteren Experiment die Sensitivität der Merkmalsenthallung gegenüber Schätzfehlern in den RIA-Parametern untersucht. Für die Simulationen wurde angenommen, dass für jede einzelne Sprachäußerung innerhalb der Testdaten jeweils unabhängige Schätzungen der Nachhallzeit und des Energieparameters vorlagen, welche durch

$$\hat{T}_{60} = T_{60} + e_{\hat{T}_{60}} \tag{6.6}$$

$$\hat{\sigma}_h^2 = \sigma_h^2 \left(1 + e_{\hat{\sigma}_h^2,\mathrm{REL}}\right) \tag{6.7}$$

gegeben waren. Dabei bezeichnet T_{60} die angenommene wahre Nachhallzeit, welche für das Büro stets zu 0,35 s und für das Wohnzimmer stets zu 0,45 s gesetzt wurde. Für die wahre Energiekonstante $\hat{\sigma}_h^2$ wurde angenommen, dass diese durch (5.171) bestimmt ist, da wie bereits erwähnt die Trainings- und die Testdaten beider Datenbanken derselben Energienormierung unterlagen. Weiterhin wurde davon ausgegangen, dass die Schätzfehler $e_{\hat{T}_{60}}$ und $e_{\hat{\sigma}_h^2,\mathrm{REL}}$ jeweils Realisierungen der beiden mittelwertfreien Zufallsvariablen $\breve{e}_{\hat{T}_{60}}$ und $\breve{e}_{\hat{\sigma}_h^2,\mathrm{REL}}$ darstellen, deren Verteilungsdichtefunktionen wie folgt definiert sind:

$$p_{\breve{e}_{\hat{T}_{60}}}\left(e_{\hat{T}_{60}}\right) := \begin{cases} c_1 \cdot \mathcal{N}\left(e_{\hat{T}_{60}}; 0, \sigma^2_{\breve{e}_{\hat{T}_{60}}}\right) & \text{für } \left|e_{\hat{T}_{60}}\right| < 2\sigma_{\breve{e}_{\hat{T}_{60}}} + 0{,}025 \\ 0 & \text{sonst} \end{cases} \tag{6.8}$$

$$p_{\breve{e}_{\hat{\sigma}_h^2,\mathrm{REL}}}\left(e_{\hat{\sigma}_h^2,\mathrm{REL}}\right) := \begin{cases} c_2 \cdot \mathcal{N}\left(e_{\hat{\sigma}_h^2,\mathrm{REL}}; 0, \sigma^2_{\breve{e}_{\hat{\sigma}_h^2,\mathrm{REL}}}\right) & \text{für } \left|e_{\hat{\sigma}_h^2,\mathrm{REL}}\right| < 2\sigma_{\breve{e}_{\hat{\sigma}_h^2,\mathrm{REL}}} \\ 0 & \text{sonst} \end{cases}. \tag{6.9}$$

Das beidseitige "Abschneiden" der GAUSS-förmigen Verteilungsdichtefunktionen sollte dabei vermeiden, dass die Schätzwerte $\hat{T}_{60}$ und $\hat{\sigma}_h^2$ negativ wurden. Die beiden positiven, reellen Normierungskonstanten c_1 und c_2 wurden dabei derart gewählt, dass das Integral über die beiden Verteilungsdichtefunktionen jeweils gleich 1 ist. Der Schätzwert der Nachhallzeit wurde zusätzlich vor der Durchführung der Merkmalsenthallung auf ganze Vielfache von 0,05 s gerundet, sodass für diese Werte der Nachhallzeit zuvor empirisch bestimmte Parameter des Beobachtungsfehlers eingesetzt werden konnten. Im Zusammenhang mit der Simulation von Schätzfehlern in der Nachhallzeit muss außerdem beachtet werden, dass bedingt durch die Erzeugung der Datenbanken die Nachhallzeit einzelner Sprachäußerungen jeweils gleichmäßig zwischen 0,3 s, 0,35 s und 0,4 s für das Büro und zwischen 0,4 s, 0,45 s und 0,5 s für das Wohnzimmer variierte, so dass bereits bei einer Standardabweichung $\sigma_{\breve{e}_{\hat{T}_{60}}} = 0$ Schätzfehler in der Nachhallzeit vorlagen.

Für die Merkmalsenthallung wurde die *IMM*-Schätzung mit $L_\mathrm{C} = 6$ eingesetzt und als A-priori-Sprachmodell dasselbe *SLDM* mit $I = 4$ Teilmodellen verwendet, das zuvor für die Untersuchungen bezüglich der Leistungsfähigkeit unterschiedlicher Modellkombinationsalgorithmen diente. Die resultierenden Wortfehlerraten in Abhängigkeit von den Standardabweichungen für die Schätzfehler in den RIA-Parametern sind jeweils für das nichtrekursive sowie das rekursive Beobachtungsmodell in Tab. 6.14 für die AURORA5-Datenbank und in Tab. 6.15 für die modifizierte AURORA4-Datenbank zusammengetragen.

Es ließ sich beobachten, dass die Auswirkungen von Schätzfehlern in den RIA-Parametern für beide Beobachtungsmodelle ähnlich waren. Die Wortfehlerrate stieg dabei für beide untersuchten Räume und Datenbanken gemittelt über alle betrachteten Werte für die Standardabweichung des Schätzfehlers im Energieparameter lediglich um etwa 10 % relativ an, wenn die Standardabweichung des Schätzfehlers in der Nachhallzeit 0,1 s betrug. Daher lässt sich schlussfolgern, dass eine zufriedenstellende Robustheit des vorgestellten Verfahrens zur Merkmalsenthallung gegenüber Schätzfehlern in den RIA-Parametern vorliegt.

6.5.3. Adaption des Erkenners auf Artefakte der Merkmalsenthallung

Im Allgemeinen lässt sich die Trajektorie der LMSK-Vektoren des sauberen Sprachsignals mit Hilfe der Merkmalsverbesserung nicht perfekt aus der Trajektorie der LMSK-Vektoren

Tabelle 6.14.: *Wortfehlerraten λ_w [%] in Abhängigkeit von den Standardabweichungen für die Schätzfehler in den RIA-Parametern für die AURORA5-Datenbank.*

		Raum							
		Büro				Wohnzimmer			
		$\sigma_{\check{e}_{\hat{\sigma}_h^2,\mathrm{REL}}}$ [dB]							
		$-\infty$	-15	-10	-5	$-\infty$	-15	-10	-5
$\sigma_{\check{e}_{\hat{T}_{60}}}$	0	**1,97**	1,97	1,97	1,97	**3,61**	3,59	3,58	3,59
	0,075	2,21	2,18	2,21	2,23	4,03	4,15	4,16	4,21
	0,1	2,34	2,22	2,41	2,35	4,21	4,22	4,27	4,35

(a) Nichtrekursives Beobachtungsmodell

		Raum							
		Büro				Wohnzimmer			
		$\sigma_{\check{e}_{\hat{\sigma}_h^2,\mathrm{REL}}}$ [dB]							
		$-\infty$	-15	-10	-5	$-\infty$	-15	-10	-5
$\sigma_{\check{e}_{\hat{T}_{60}}}$	0	**2,17**	2,14	2,13	2,15	**4,13**	4,16	4,17	4,24
	0,075	2,30	2,32	2,30	2,39	4,25	4,18	4,22	4,32
	0,1	2,29	2,40	2,38	2,35	4,31	4,37	4,43	4,47

(b) Rekursives Beobachtungsmodell

Tabelle 6.15.: *Wortfehlerraten λ_w [%] in Abhängigkeit von den Standardabweichungen für die Schätzfehler in den RIA-Parametern für die AURORA4-Datenbank.*

		Raum							
		Büro				Wohnzimmer			
		$\sigma_{\check{e}_{\hat{\sigma}_h^2,\mathrm{REL}}}$ [dB]							
		$-\infty$	-15	-10	-5	$-\infty$	-15	-10	-5
$\sigma_{\check{e}_{\hat{T}_{60}}}$	0	**27,77**	27,55	27,44	27,44	**40,44**	40,52	40,70	41,07
	0,075	29,17	29,13	29,28	29,50	42,65	42,95	41,03	41,84
	0,1	29,94	29,98	33,30	28,84	41,99	44,09	44,35	43,68

(a) Nichtrekursives Beobachtungsmodell

		Raum							
		Büro				Wohnzimmer			
		$\sigma_{\check{e}_{\hat{\sigma}_h^2,\mathrm{REL}}}$ [dB]							
		$-\infty$	-15	-10	-5	$-\infty$	-15	-10	-5
$\sigma_{\check{e}_{\hat{T}_{60}}}$	0	**26,22**	26,85	26,92	27,22	**40,18**	41,62	41,36	41,73
	0,075	28,55	28,21	28,21	28,66	42,69	41.73	41.92	41.77
	0,1	28,40	29.43	29.69	29.58	43,68	42.84	43.54	43.31

(b) Rekursives Beobachtungsmodell

des verhallten Sprachsignals rekonstruieren. Die verbleibenden Artefakte führen dann zu einer Veränderung der statistischen Eigenschaften der Trajektorien gegenüber dem Fall in Abwesenheit von Nachhall. Beruhend auf dieser Diskrepanz zwischen Test- und Trainingsbedingungen steigt die Wortfehlerrate bei der Spracherkennung gewöhnlich an. Eine Möglichkeit diesem Problem zu begegnen besteht in einer sinnvollen Anpassung der Trainingsbedingungen. Dies lässt sich bewerkstelligen, indem die sauberen Trainingssprachsignale zunächst umgebungsspezifisch künstlich verhallt und anschließend mit Hilfe der Merkmalsverbesserung wieder enthallt werden. Die Erzeugung der künstlichen RIAs kann dabei mit Hilfe der Spiegelquellenmethode auf dieselbe Art und Weise wie in Kap. 6.4 geschehen.

In Tab. 6.16 sind die auf diese Weise erzielten Fehlerraten für die AURORA5- und die modifizierte AURORA4-Datenbank zusammengetragen. Zusätzlich sind in derselben Tabel-

Tabelle 6.16.: *Fehlerraten [%] für ausgewählte Kombinationen von unterschiedlichen Trainingsbedingungen und der An- bzw. Abwesenheit der Merkmalsverbesserung.*

Konditionen	Raum Büro	Raum Wohnzimmer
	λ_w	
Training auf sauberen Sprachsignalen, Merkmalsverbesserung vor der Erkennung	1,97	3,61
Training auf enthallten Sprachsignalen, Merkmalsverbesserung vor der Erkennung	**2,00**	**3,35**
Training auf verhallten Sprachsignalen, Erkennung ohne Merkmalsverbesserung	1,29	2,61

(a) AURORA5-Datenbank

Konditionen	Raum Büro				Raum Wohnzimmer			
	λ_{Subst}	λ_{Ausl}	λ_{Einf}	λ_w	λ_{Subst}	λ_{Ausl}	λ_{Einf}	λ_w
Training auf sauberen Sprachsignalen, Merkmalsverbesserung vor der Erkennung	20,07	2,54	5,16	**27,77**	29,10	3,61	7,73	**40,44**
Training auf enthallten Sprachsignalen, Merkmalsverbesserung vor der Erkennung	**17,46**	**2,43**	**4,49**	**24,38**	**23,72**	**3,61**	**5,64**	**32,97**
Training auf verhallten Sprachsignalen, Erkennung ohne Merkmalsverbesserung	18,01	3,06	3,17	**24,24**	26,26	6,08	3,98	**36,32**

(b) Modifizierte AURORA4-Datenbank

le die Resultate für das Training des Erkenners mit verhallten Sprachsignalen, wie sie bereits auch in Tab. 6.3b und Tab. 6.4b aufgeführt sind, sowie die Resultate für die zuvor betrachtete Merkmalsenthallung in Kombination mit einem Training des Spracherkenners auf sauberen Sprachäußerungen aus Tab. 6.5, Tab. 6.8a und Tab. 6.9a gegenübergestellt.

Für die AURORA5-Datenbank ließ sich beobachten, dass das Training des Spracherkenners auf Artefakte nach der Merkmalsverbesserung lediglich für das Wohnzimmer zu einer geringfügigen Abnahme der Wortfehlerrate führte. Bedauerlicher ließ sich damit nicht die Leistungsfähigkeit erreichen, die beim Training des Erkenners mit verhallten Sprachsignalen erzielt werden konnte. Hingegen nahm für die modifizierte AURORA4-Datenbank die Wortfehlerrate bedingt durch das Training des Spracherkenners auf Artefakte nach der Merkmalsverbesserung deutlich stärker ab, insbesondere für das Wohnzimmer. Während für das Büro eine ähnliche Leistungsfähigkeit wie beim Training des Erkenners mit verhallten Sprachsignalen erreicht werden konnte, wurde diese für das Wohnzimmer sogar übertroffen.

Eine mögliche Erklärung könnte darin bestehen, dass im Falle eines sehr umfangreichen Vokabulars in Kombination mit einer hohen Nachhallzeit T_{60} der Effekt des Nachhalls aus verhallten Trainingssprachsignalen schlechter gelernt werden kann, da der zu berücksichtigende links- bzw. rechtsseitige Kontext zu einem HMM-Zustand deutlich stärker variieren kann. Dieses Problem tritt beim Training mit enthallten Trainingsdaten in deutlich geringerem Maße auf, da durch die Enthallung die zeitliche Dispersion reduziert und damit der links- und rechtsseitige Kontext teilweise eingeschränkt wird. Ein weitere Ursache für den stärkeren Effekt des Trainings der Parameter des akustischen Modells auf die nach der Merkmalsverbesserung verbleibenden Artefakte bei der modifizierten AURORA4-Datenbank besteht darin, dass die Auswirkung dieser Artefakte an sich im Vergleich zur AURORA5-Datenbank größer ist, weil zwischen einer größeren Anzahl an Wörtern unterschieden werden muss.

Zusammenfassend lässt sich feststellen, dass sich mit einem derartigen kombinierten Ansatz auf der AURORA5-Datenbank etwa 80 % und auf der AURORA4-Datenbank etwa 70 % der Fehler, die durch den Nachhall entstanden sind, beheben ließen.

6.6. Ergebnisse zur gemeinsamen Merkmalsenthallung und -entstörung

In einem letzten Experiment wurde die Leistungsfähigkeit des vorgestellten Verfahrens in Gegenwart von sowohl Nachhall als auch Hintergrundstörungen untersucht. Als A-priori-Sprachmodell wurde für jede Datenbank dasjenige *SLDM* ausgewählt, welches bereits bei den Experimenten zur Merkmalsenthallung die niedrigste Wortfehlerrate lieferte. Die Parameter des A-priori-Modells für die Störung, d.h. der Mittelwertvektor $\boldsymbol{\mu}_{\mathbf{n}}$ und die Kovarianzmatrix $\boldsymbol{\Sigma}_{\mathbf{n}}$ (siehe Kap. 5.1.2), wurden empirisch unter Verwendung der jeweils 15 ersten und letzten Segmente einer Sprachäußerung auf der AURORA5-Datenbank bestimmt, da in den entsprechenden Zeiträumen keine Sprachaktivität vorlag. Bei der modifizierten AURORA4-Datenbank wurden für diesen Zweck die 50 ersten und letzten Segmente einer Sprachäußerung verwendet. Die Merkmalsverbesserung wurde jeweils mit Hilfe des nichtrekursiven und des rekursiven Beobachtungsmodells durchgeführt, wobei $L_{\mathrm{C}} = L_{\mathrm{R}} = 6$ angenommen wurde. Die bei der Spracherkennung erzielten Wortfehlerraten sind in Tab. 6.17 für die AURORA5-Datenbank und in Tab. 6.18 für die modifizierte AURORA4-Datenbank aufgeführt. Aus den Ergebnissen lässt sich eine leichte Tendenz zugunsten des nichtrekursiven Beobachtungs-

Tabelle 6.17.: *Wortfehlerraten λ_w [%] für die AURORA5-Datenbank erzielt mit der gemeinsamen Merkmalsenthallung und -entstörung.*

		Raum	
		Büro	Wohnzimmer
SNR [dB]	15	7,47	12,21
	10	16,83	24,04
	5	35,13	44,33
	0	62,44	69,51

(a) Nichtrekursives Beobachtungsmodell

		Raum	
		Büro	Wohnzimmer
SNR [dB]	15	7,77	12,54
	10	17,27	24,62
	5	35,67	44,70
	0	62,93	70,77

(b) Rekursives Beobachtungsmodell

Tabelle 6.18.: *Fehlerraten [%] für die modifizierte AURORA4-Datenbank erzielt mit der gemeinsamen Merkmalsenthallung und -entstörung.*

		Raum							
		Büro				Wohnzimmer			
		λ_{Subst}	λ_{Ausl}	λ_{Einf}	λ_w	λ_{Subst}	λ_{Ausl}	λ_{Einf}	λ_w
SNR [dB]	15	31,09	4,53	10,64	**46,26**	43,76	5,45	11,57	**60,77**
	10	46,15	7,29	12,63	**66,08**	55,32	10,28	14,11	**79,71**
	5	62,10	14,22	10,31	**86,63**	65,60	16,87	9,65	**92,12**
	0	61,80	28,88	4,75	**95,43**	62,14	31,68	3,28	**97,09**

(a) Nichtrekursives Beobachtungsmodell

		Raum							
		Büro				Wohnzimmer			
		λ_{Subst}	λ_{Ausl}	λ_{Einf}	λ_w	λ_{Subst}	λ_{Ausl}	λ_{Einf}	λ_w
SNR [dB]	15	31,45	4,71	10,94	**47,11**	42,36	5,67	11,23	**59,26**
	10	46,30	7,26	11,93	**65,49**	56,17	9,80	11,68	**77,64**
	5	61,55	14,62	9,39	**85,56**	64,86	17,13	8,80	**90,79**
	0	58,05	32,52	4,38	**94,95**	57,46	36,83	2,84	**97,13**

(b) Rekursives Beobachtungsmodell

modells auf der AURORA5-Datenbank erkennen, die jedoch auf der AURORA4-Datenbank ins Gegenteil umschlägt. Die Unterschiede der Ergebnisse sind jedoch nur gering, so dass auf keinen sinnvollen Vorteil eines bestimmten Beobachtungsmodells anhand dieser Simulationsergebnisse geschlossen werden sollte.

Bei einem Vergleich der Ergebnisse mit denen des gewöhnlichen *ETSI-SFE* ohne nachgeschaltete Merkmalsverbesserung in Tab. 6.1 und Tab. 6.2 fällt auf, dass die Leistungsfähigkeit deutlich mit sinkendem *SNR* abnahm. Konnten bei einem *SNR* von 15 dB bei der AURORA5-Datenbank immerhin noch etwa 65 % der durch den Nachhall und die Hintergrundstörungen verursachten Fehler behoben werden, so waren es bei einem *SNR* von 0 dB nur noch maximal etwa 30 %. Ein ähnliches Verhalten zeigte sich auch bei der modifizierten AURORA4-Datenbank, wobei die erzielten Verbesserungen im Vergleich zur AURORA5-Datenbank insgesamt deutlich geringer waren. Während bei einem *SNR* von 15 dB noch etwa 40 % bzw. 53 % der Fehler beim Büro bzw. Wohnzimmer korrigiert werden konnten, betrug der Anteil korrigierter Fehler bei einem *SNR* von 0 dB nur noch etwa 5 % beim Büro und etwa 2 % beim Wohnzimmer. Die schlechtere Leistungsfähigkeit der Merkmalsverbesserung auf der AURORA4-Datenbank hängt auch hier mit dem deutlich größeren Vokabular zusammen, wodurch nach der Verbesserung verbleibende Artefakte vom Spracherkenner eher falsch interpretiert werden können.

Die abnehmende Leistungsfähigkeit bei sinkenden Werten des *SNR* besitzt hauptsächlich zwei Ursachen. Zum einen ist das zur Beschreibung der Hintergrundstörung verwendete A-priori-Modell nur bedingt geeignet, da die Störungen, welche zur Erzeugung beider untersuchter Datenbanken herangezogen wurden, einen besonders instationären Charakter besitzen. Zum anderen sind, wie bereits in Kap. 6.4 angemerkt, das Modell des Beobachtungsfehlers sowie dessen Parameter bei Vorhandensein von Störungen stark abhängig vom lokalen *SNR*. In den Experimenten in dieser Arbeit wurde diese Tatsache im Sinne einer Vereinfachung nicht berücksichtigt, wodurch jedoch starke Einbußen in der Leistungsfähigkeit der Merkmalsverbesserung hingenommen werden mussten.

Trotz dieser beiden starken Vereinfachungen übertraf die Leistungsfähigkeit des in dieser Arbeit vorgeschlagenen Verfahrens deutlich jene des *ETSI-AFE* für hohe Werte des *SNR*, wie durch einen Vergleich der Wortfehlerraten in Tab. 6.17 bzw. Tab. 6.18 mit den in Tab. 6.3 und Tab. 6.4 ersichtlich wird. Erst für sehr niedrige *SNR*-Werte von 5 dB bzw. 0 dB lieferte das speziell zur Entstörung konzipierte *ETSI-AFE* bessere Ergebnisse.

Obwohl das Training des Spracherkenners mit verhallten Sprachsignalen nicht zur Robustheit gegenüber Hintergrundstörungen beiträgt, waren die damit erzielten Wortfehlerraten auf der AURORA4-Datenbank durchgehend besser als diejenigen, welche mit der Merkmalsverbesserung oder mit dem *ETSI-AFE* erhalten worden sind. Im Gegenteil dazu war jedoch die Merkmalsverbesserung auf der AURORA5-Datenbank bei Vorhandensein von Hintergrundstörungen im Vergleich dazu stets dominant.

Aufgrund der sehr groben Modellierung der Störung durch das A-priori-Modell sowie der groben Modellierung des Beobachtungsfehlers wird davon ausgegangen, dass die Leistungsfähigkeit des vorgestellten Verfahrens durch die Verbesserung beider Modelle beträchtlich gesteigert werden kann. Diese Aspekte bleiben jedoch der zukünftigen Forschung vorbehalten und werden in dieser Arbeit nicht weiter behandelt.

7. Zusammenfassung und Ausblick

Im Rahmen dieser Arbeit wurde ein Verfahren zur Verbesserung akustischer Merkmale im Hinblick auf eine robuste Spracherkennung in Gegenwart von Nachhall und Hintergrundstörungen entwickelt, wobei der Schwerpunkt auf der Kompensation des Nachhalls lag. Als akustische Merkmale wurden dabei die log-MEL-spektralen Merkmale betrachtet, da sie die unmittelbare Vorstufe zur Berechnung der *MFCC*s darstellen. Es ist dabei besonders zu betonen, dass aufgrund der weiten Verbreitung der *MFCC*s in Verbindung mit der dabei erzielten hohen Erkennungsleistung eine hohe Relevanz des hier vorgestellten Ansatzes gegeben ist. Die hohe Relevanz wird zudem unterstützt durch eine hohe Flexibilität einer jeden merkmalsbasierten Methode, da diese prinzipiell beliebige Strukturen des Spracherkenners zulässt.

Das Konzept der auf BAYES'scher Inferenz basierenden Merkmalsverbesserung wurde in Kap. 5 vorgestellt. Es nutzt die Information von A-priori-Modellen der Sprache und der Hintergrundstörung sowie eines Beobachtungsmodells in einer statistisch optimalen Art. Dabei wurden zur Beschreibung des A-priori-Wissens über die Merkmalsvektortrajektorie der sauberen Sprache schaltende, lineare dynamische Modelle eingesetzt, wobei insbesondere auch Modelle höherer Ordnung als eins in Betracht gezogen wurden. Die entsprechenden Modellparameter wurden mit Hilfe des *EM*-Algorithmus und einer Menge von Trainingssprachäußerungen bestimmt. Als Folge dessen, dass der *EM*-Algorithmus an sich eher ein Prinzip als einen konkreten Algorithmus darstellt, wurden dafür zunächst die notwendigen Schätzformeln für sämtliche *SLDM*-Parameter hergeleitet. Da es sich beim *EM*-Algorithmus um ein iteratives Verfahren handelt, werden zu seiner Anwendung Startwerte für die *SLDM*-Parameter benötigt. Zu diesem Zweck wurde eine neuartige, stochastische Initialisierungsmethode vorgeschlagen, deren Prinzip ähnlich dem des K-MEANS++-Algorithmus ist.

Ein weiterer besonderer Aspekt, der in Kap. 5 behandelt wurde, liegt in der Herleitung des Beobachtungsmodells zur Beschreibung des Zusammenhanges zwischen den log-MEL-spektralen Merkmalen des verhallten und gestörten Sprachsignals sowie den log-MEL-spektralen Merkmalen des sauberen Sprachsignals und des Störsignals. Dieser Zusammenhang wurde zunächst auf der Grundlage der RIA zwischen dem Sprecher und dem Mikrophon, welche die Mehrwegeausbreitung des Signals kennzeichnet, hergeleitet. Um eine in der Regel hoch sensible, blinde Schätzung der in praktisch relevanten Anwendungen gewöhnlich unbekannten und zudem zeitvarianten RIA zu vermeiden, wurde diese durch ein statistisches Modell beschrieben. Das verwendete Modell besitzt nur zwei Parameter, welche die Energie und das Abklingverhalten der RIA charakterisieren. Die beiden Parameter können deutlich einfacher und robuster als die vollständige RIA blind aus dem eingehenden Mikrophonsignal geschätzt werden. In der Arbeit wurde nun vorgeschlagen, wie auf der Basis des statistischen Modells der RIA lediglich unter Verwendung der Modellparameter ein sinnvolles Beobachtungsmodells berechnet werden kann.

Aufgrund des dispersiven Effektes hängt ein log-MEL-spektrales Merkmal eines verhallten Sprachsignals gewöhnlich von mehreren, zeitlich zurückliegenden log-MEL-spektralen Merkmalen des zugehörigen sauberen Sprachsignals ab, welche bei der Auswertung der Beobachtungsfunktion berücksichtigt werden müssen. Zur Reduktion des damit im Zusammenhang stehenden Rechen- und Speicheraufwandes wurde ein rekursives Beobachtungsmodell hergeleitet, wobei die Rekursionslänge vollkommen variabel gewählt werden kann.

Sowohl für das nicht rekursive als auch das rekursive Beobachtungsmodell wurde der Beobachtungsfehler stark vereinfacht als eine Realisierung eines weißen, GAUSS'schen Zufallsprozesses beschrieben. Obwohl die Unabhängigkeit einzelner zeitlich aufeinander folgender Fehler bei Weitem nicht gegeben ist, konnte später in Kap. 6 jedoch zumindest mit Hilfe von Merkmalen realer Sprachsignale experimentell gezeigt werden, dass das Histogramm des Beobachtungsfehlers eine annähernd GAUSS-glockenförmige Gestalt aufweist.

Zur praktischen Umsetzung der Inferenz wurden in dieser Arbeit suboptimale Modellkombinationsalgorithmen verwendet, um einen zeitlich konstant bleibenden Aufwand an Stelle eines exponentiell wachsenden zu erzielen. Das Prinzip der dabei kombinierten, teilmodellspezifischen Inferenzen beruhte auf dem eines erweiterten iterativen KALMAN-Filters.

In Kap. 6 wurden experimentelle Untersuchungen zum Verfahren der Merkmalsverbesserung durchgeführt. Dazu wurden zwei verschiedene Sprachdatenbanken, die AURORA5-Datenbank und eine modifizierte Version der AURORA4-Datenbank, herangezogen. Während die AURORA5-Datenbank unter anderem Sprachäußerungen von einzelnen Ziffern und Ziffernketten beinhaltet, sind in der AURORA4-Datenbank Äußerungen kontinuierlich gesprochener Sprache in Form ganzer Sätze enthalten. Beide Datenbanken bestehen aus sauberen Sprachsignalen und deren künstlich erzeugten verhallten sowie verhallten und zusätzlich gestörten Versionen. Zur künstlichen Verhallung wurden für beide Datenbanken zwei unterschiedliche virtuelle Räume mit Nachhallzeiten von jeweils etwa 0,35 s und 0,45 s angenommen. Als additive Hintergrundstörungen wurden Ausschnitte aus Aufnahmen aus typischen Innenräumen benutzt, um realistisch Störungen nachzuahmen.

Die Leistungsfähigkeit der Merkmalsverbesserung wurde in dieser Arbeit indirekt über die nach einer Spracherkennung erzielte Wortfehlerrate bewertet. Es wurden dabei Untersuchungen sowohl zur ausschließlichen Merkmalsenthallung als auch zur gemeinsamen Entstörung und Enthallung von Merkmalen durchgeführt.

Die Ergebnisse zur Merkmalsenthallung zeigen eine deutliche Reduktion der Wortfehlerrate für alle drei betrachteten Modellkombinationsalgorithmen gegenüber dem Fall ohne der Verwendung jeglicher Merkmalsenthallung. Insbesondere steigerte sich die Leistungsfähigkeit bei einer Berücksichtigung eines gewissen zeitlichen Kontexts aus der Zukunft, was durch eine geeignete Erweiterung des Zustandsvektors bei der erweiterten KALMAN-Filterung erreicht wurde. Bemerkenswert im Bezug auf die Wahl des A-priori-Modells zur Beschreibung der Sprache im Merkmalsbereich ist die Tatsache, dass bereits mit einem einzigen linearen dynamischen Modell die Wortfehler, die durch den Nachhall verursacht worden sind, um bis zu 75 % für die Ziffernkettenerkennung reduziert werden konnten. Durch eine moderate Vergrößerung der Anzahl der Teilmodelle des *SLDM* auf 4 ließ sich der Prozentanteil auf bis zu 80 % erhöhen. Bei der Erkennung von kontinuierlich gesprochener Sprache mit großem Vokabular betrug dieser Prozentanteil immerhin noch etwa 50 %, da die nach der Verbesserung verbliebenen Fehler tendenziell schwerwiegendere Auswirkungen bedingt durch die erhöhte Komplexität der Erkennungsaufgabe hatten.

Bezüglich der Wahl des A-priori-Modells der Sprache konnte weiterhin beobachtet wer-

den, dass die vorgeschlagene Methode der Initalisierung der *SLDM*-Parameter unter der Annahme einer geeigneten Wahl der Anzahl der *EM*-Iterationen letzendlich zu einer geringfügig verringerten Wortfehlerrate führen kann. Jedoch war der erzielte Gewinn nur minimal, was nach Ansicht des Autors mit dem verwendeten Kriterium zum Training der *SLDM*-Parameter, nämlich der Maximierung der Likelihood der Trainingsdaten, zusammenhängt, welches nicht unmittelbar mit dem der Minimierung der Wortfehlerrate zusammenhängt. Es ist zu vermuten, dass durch ein geeigneteres Kriterium, welches zusätzlich eine zeitlich lokale und getrennte Aktivität einzelner Teilmodelle des *SLDM* fordert, die vorgeschlage Initialisierungsmethode an Bedeutung gewinnen wird.

Die Erhöhung der *SLDM*-Ordnung bewirkte nur bei der Nutzung eines einzigen linearen, dynamischen Modells als A-priori-Modell der Sprache eine Reduktion der Wortfehlerrate, wobei der Anteil der Verbesserung mit der Erhöhung der Ordnung abnahm. In diesem besonderen Fall ist keine Anwendung des *EM*-Algorithmus zum *SLDM*-Training notwendig, da ein analytischer Ausdruck zur direkten Berechnung der Parameter existiert. Zudem ist bei einem einzigen Teilmodell keine suboptimale Modellkombination zur approximativen Umsetzung der Inferenz erforderlich. Beide Aspekte können als Ursache dafür angesehen werden, dass durch die Verwendung mehrerer Teilmodelle des *SLDM* keine Verringerung der Wortfehlerrate erreicht werden konnte.

Bezüglich des Beobachtungsmodells konnte einerseits experimentell festgestellt werden, dass mit dem rekursiven Beobachtungsmodell ähnliche Wortfehlerraten erzielt werden konnten, wobei sowohl der Rechen- als auch Speicheraufwand im Vergleich zum nicht rekursiven Beobachtungsmodell geringfügig reduziert werden konnten. Andererseits konnte eine gewisse Robustheit der Merkmalsverbesserung gegenüber Schätzfehlern in den beiden Parametern der RIA experimentell festgestellt werden. Unter der Annahme eines annähernd GAUSS-verteilten Schätzfehlers in der Nachhallzeit stieg die Wortfehlerrate beispielsweise lediglich um 10 % an, wenn die Standardabweichung des Schätzfehlers approximativ 0,1 s betrug.

Um den Erkenner auf nach der Merkmalsverbesserung noch vorhandene Artefakte anzupassen, wurden überdies Experimente durchgeführt, bei dem die für das Training des Spracherkenners verwendeten Sprachsignale vorab künstlich verhallt und anschließend auf Merkmalsebene wieder enthallt wurden, bevor das Training der akustischen Modelle des Spracherkenners erfolgte. Dieses Vorgehen zeigte den größten Effekt bei der Verwendung der Sprachdatenbank mit großem Vokabular, d. h. der AURORA4-Datenbank. Ein möglicher Grund dafür könnte darin bestehen, dass sich Artefakte stärker auswirken, wenn zwischen einer größeren Anzahl an Wörtern bei der Erkennung unterschieden werden muss. Bei der AURORA5-Datenbank war deshalb nur eine geringe Wirkung zu beobachten.

In den abschließenden Experimenten zur gemeinsamen Enthallung und Entstörung der akustischen Merkmale lieferten das rekursive und das nicht rekursive Beobachtungsmodell sehr ähnliche Ergebnisse. Zusammenfassend lässt sich festhalten, dass sich die Leistungsfähigkeit der Merkmalsverbesserung mit sinkendem *SNR* deutlich verringerte. Konnten bei der Ziffernkettenerkennung bei einem *SNR* von 15 dB noch etwa 65 % der durch den Nachhahll und die Hintergrundstörungen eingeführten Fehler behoben werden, waren es bei einem *SNR* von 0 dB nur noch etwa 30 %. Bei der Erkennung kontinuierlicher Sprache waren es dagegen maximal 53 % bei einem *SNR* von 15 dB und nur noch maximal 5 % bei einem *SNR* von 0 dB. Dieser Effekt besitzt hauptsächlich zwei Ursachen.

Zum einen ist das verwendete A-priori-Modell zur Beschreibung der Charakteristik der

Störung relativ grob in der Hinsicht, als dass es nur stationäre Störungen vernünftig erfassen kann. Da die verwendeten Signale der Hintergrundstörung jedoch einen vorwiegend instationären Charakter aufwiesen, war die Modellierung der Störung überaus ungenau, was sich am meisten bei niedrigen Werten des *SNR* bemerkbar machte. Als Ausblick in diesem Zusammenhang ist eine Verbesserung des A-priori-Modells der Störung zu nennen, wovon anzunehmen ist, dass dies deutlich zur Verbesserung der Leistungsfähigkeit der Merkmalsverbesserung beitragen kann.

Zum anderen ist die Modellierung des Beobachtungsfehlers in Gegenwart von Hintergrundstörungen unzureichend, da dabei die Hintergrundstörung vollständig ignoriert wird. Eine Möglichkeit der Verbesserung besteht in der Annahme eines Modells mit zeitvarianten Parametern, deren Wahl beispielsweise abhängig von einer Schätzung des *SNR* gemacht werden könnte.

Trotz beider Defizite übertraf die Leistungsfähigkeit des vorgestellten Verfahrens zur gemeinsamen Enthallung und Entstörung akustischer Merkmale bei hohen Werten des *SNR* die des *ETSI-AFE*, welches ein renommiertes Verfahren zur Merkmalsentstörung darstellt.

A. Anhang

A.1. Herleitung des *EM*-Algorithmus zum Training von *SLDM*s beliebiger Ordnung

In diesem Abschnitt werden die Rekursionsgleichungen zur Schätzung der *SLDM*-Parameter

$$\theta = \left\{ \boldsymbol{\mu}_{\mathbf{x},i}, \boldsymbol{\Sigma}_{\mathbf{x},i}, \mathbf{A}_{i,\nu}, \mathbf{b}_i, \mathbf{V}_i, \psi_i, a_{i,k} \middle| i,k \in \{1,\ldots,I\}, \nu \in \{1,\ldots,L_{\mathrm{AR}}\} \right\} \tag{A.1}$$

mit Hilfe von Trainingsdaten in Form einer Menge von unabhängigen Merkmalsvektorsequenzen $\mathfrak{X}$ gemäß dem *EM*-Algorithmus hergeleitet. Dabei wird ausgehend von einer initialen Parametermenge $\theta^{\{0\}}$ iterativ eine Folge von Parametermengen $\left\{ \theta^{\{l\}} \middle| l \in \mathbb{N} \right\}$ bestimmt. Die Berechnung der Menge $\theta^{\{l+1\}}$ vollzieht sich in zwei Teilschritten, dem *Expectation*- und dem *Maximization*-Schritt, welche dem Algorithmus seinen Namen geben und im Folgenden detailliert beschrieben werden. Die Herleitung ist stark angelehnt an diejenige in [Mur98], wo jedoch nur der Fall der Modellordnung $L_{\mathrm{AR}} = 1$ behandelt wird.

A.1.1. *Expectation*-Schritt

Im ersten Schritt wird der Erwartungswert der Loglikelihood der kompletten Daten bestehend aus der Menge der Merkmalsvektorsequenzen $\mathfrak{X}$ und der Menge der zugehörigen, nicht beobachtbaren Zustandssequenzen $\mathfrak{Z}$ bedingt auf $\mathfrak{X}$ und die zuvor berechnete Parametermenge $\theta^{\{l\}}$ gemäß

$$\mathscr{Q}_{l+1}(\theta) := \mathrm{E}\left[\ln\left\{ p_{\check{\mathfrak{X}},\check{\mathfrak{Z}}}(\mathfrak{X},\mathfrak{Z}) \right\} \middle| \mathfrak{X}; \theta^{\{l\}} \right] \tag{A.2}$$

$$= \sum_{\{\mathfrak{Z}\}} \ln\{p(\mathfrak{X},\mathfrak{Z})\} P\left(\mathfrak{Z} \middle| \mathfrak{X}; \theta^{\{l\}} \right). \tag{A.3}$$

berechnet, wobei in (A.3) die Summation als Summation über alle möglichen Realisierungen $\mathfrak{Z}$ zu verstehen ist und im Sinne der Lesbarkeit die Indizes der Verteilungsdichtefunktionen und der Wahrscheinlichkeitsmassefunktionen weggelassen wurden. Dabei wird der Erwartungswert gebildet, um die Abhängigkeit der Loglikelihood von der nicht beobachtbaren und daher unbekannten Menge der Zustandssequenzen $\mathfrak{Z}$ zu eliminieren.

Unter Ausnutzung der Unabhängigkeit der Sprachäußerungen sowie der Definition des

SLDM in (5.13) lässt sich die Loglikelihood der kompletten Daten gemäß

$$\ln\{p(\mathfrak{X},\mathfrak{Z})\} = \sum_{n=1}^{N} \ln\left\{p\left(\mathbf{x}_{1:M_n}^{(n)}, \zeta_{1:M_n}^{(n)}\right)\right\} \tag{A.4}$$

$$= \sum_{n=1}^{N}\left\{\sum_{m=1}^{L_{\mathrm{AR}}}\left[\ln\left\{p\left(\mathbf{x}_m^{(n)}\middle|\,\zeta_m^{(n)}\right)\right\} + \ln\left\{P\left(\zeta_m^{(n)}\right)\right\}\right]\right.$$
$$\left. + \sum_{m=L_{\mathrm{AR}}+1}^{M_n}\left[\ln\left\{p\left(\mathbf{x}_m^{(n)}\middle|\,\mathbf{x}_{m-L_{\mathrm{AR}}:m-1}^{(n)}, \zeta_m^{(n)}\right)\right\} + \ln\left\{P\left(\zeta_m^{(n)}\middle|\,\zeta_{m-1}^{(n)}\right)\right\}\right]\right\} \tag{A.5}$$

ausdrücken. Der Erwartungswert der Loglikelihood (A.2) kann damit unter Verwendung von (A.5) und der bereits in (5.26) und (5.27) definierten bedingten Zustandswahrscheinlichkeiten

$$\eta_m^{(n,l)}(i) = P\left(\zeta_m^{(n)} = i\middle|\mathbf{x}_{1:M_n}^{(n)}; \theta^{\{l\}}\right) \tag{A.6}$$

$$\xi_m^{(n,l)}(k,i) = P\left(\zeta_m^{(n)} = i, \zeta_{m-1}^{(n)} = k\middle|\mathbf{x}_{1:M_n}^{(n)}; \theta^{\{l\}}\right) \tag{A.7}$$

gemäß

$$\mathscr{Q}_{l+1}(\theta)$$

$$= \sum_{n=1}^{N}\sum_{\left\{\zeta_{1:M_n}^{(n)}\right\}} \ln\left\{p\left(\mathbf{x}_{1:M_n}^{(n)}, \zeta_{1:M_n}^{(n)}\right)\right\} P\left(\zeta_{1:M_n}^{(n)}\middle|\,\mathbf{x}_{1:M_n}^{(n)}; \theta^{\{l\}}\right) \tag{A.8}$$

$$= \sum_{n=1}^{N}\left\{\sum_{m=1}^{L_{\mathrm{AR}}}\sum_{i=1}^{I} P\left(\zeta_m^{(n)} = i\middle|\,\mathbf{x}_{1:M_n}^{(n)}; \theta^{\{l\}}\right)\left[\ln\left\{p\left(\mathbf{x}_m^{(n)}\middle|\,\zeta_m^{(n)} = i\right)\right\} + \ln\left\{P\left(\zeta_m^{(n)} = i\right)\right\}\right]\right.$$
$$+ \sum_{m=L_{\mathrm{AR}}+1}^{M_n}\sum_{i=1}^{I}\left[P\left(\zeta_m^{(n)} = i\middle|\mathbf{x}_{1:M_n}^{(n)}; \theta^{\{l\}}\right)\ln\left\{p\left(\mathbf{x}_m^{(n)}\middle|\,\mathbf{x}_{m-L_{\mathrm{AR}}:m-1}^{(n)}, \zeta_m^{(n)} = i\right)\right\}\right.$$
$$\left.\left. + \sum_{k=1}^{I} P\left(\zeta_m^{(n)} = i, \zeta_{m-1}^{(n)} = k\middle|\mathbf{x}_{1:M_n}^{(n)}; \theta^{\{l\}}\right)\ln\left\{P\left(\zeta_m^{(n)} = i\middle|\,\zeta_{m-1}^{(n)} = k\right)\right\}\right]\right\}. \tag{A.9}$$

$$= \sum_{n=1}^{N}\left\{\sum_{m=1}^{L_{\mathrm{AR}}}\sum_{i=1}^{I}\eta_m^{(n,l)}(i)\left[\ln\left\{p\left(\mathbf{x}_m^{(n)}\middle|\,\zeta_m^{(n)} = i\right)\right\} + \ln\{\psi_i\}\right]\right.$$
$$+ \sum_{m=L_{\mathrm{AR}}+1}^{M_n}\sum_{i=1}^{I}\left[\eta_m^{(n,l)}(i)\ln\left\{p\left(\mathbf{x}_m^{(n)}\middle|\,\mathbf{x}_{m-L_{\mathrm{AR}}:m-1}^{(n)}, \zeta_m^{(n)} = i\right)\right\}\right.$$
$$\left.\left.\left. + \sum_{k=1}^{I}\xi_m^{(n,l)}(k,i)\ln\{a_{k,i}\}\right]\right\}. \tag{A.10}$$

formuliert werden. Ersetzt man in einem letzten Schritt noch die verbleibenden Verteilungsdichtefunktionen $p\left(\mathbf{x}_m^{(n)}\middle|\,\zeta_m^{(n)} = i\right)$ und $p\left(\mathbf{x}_m^{(n)}\middle|\,\mathbf{x}_{m-L_{\mathrm{AR}}:m-1}^{(n)}, \zeta_m^{(n)} = i\right)$ durch die gemäß der

Definition des *SLDM* gegebenen Ausdrücke in (5.13), so erhält man das endgültige Resultat

$$
\begin{aligned}
&\mathscr{Q}_{l+1}(\theta) \\
&= \sum_{n=1}^{N} \Bigg\{ \sum_{m=1}^{L_{\mathrm{AR}}} \sum_{i=1}^{I} \eta_m^{(n,l)}(i) \left[\ln\left\{ \mathscr{N}\left(\mathbf{x}_m^{(n)}; \boldsymbol{\mu}_{\mathbf{x},i}, \boldsymbol{\Sigma}_{\mathbf{x},i} \right) \right\} + \ln\{\psi_i\} \right] \\
&\qquad + \sum_{m=L_{\mathrm{AR}}+1}^{M_n} \sum_{i=1}^{I} \Bigg[\eta_m^{(n,l)}(i) \ln\left\{ \mathscr{N}\left(\mathbf{x}_m^{(n)}; \sum_{\nu=1}^{L_{\mathrm{AR}}} \mathbf{A}_{i,\nu} \mathbf{x}_{m-\nu}^{(n)} + \mathbf{b}_i, \mathbf{V}_i \right) \right\} \\
&\qquad\qquad + \sum_{k=1}^{I} \xi_m^{(n,l)}(k,i) \ln\{a_{k,i}\} \Bigg] \Bigg\} \qquad \text{(A.11)} \\
&= \sum_{n=1}^{N} \Bigg\{ \sum_{m=1}^{L_{\mathrm{AR}}} \sum_{i=1}^{I} \eta_m^{(n,l)}(i) \Bigg[\left(-\frac{1}{2} \right) \Bigg(\left(\mathbf{x}_m^{(n)} - \boldsymbol{\mu}_{\mathbf{x},i} \right)^T \boldsymbol{\Sigma}_{\mathbf{x},i}^{-1} \left(\mathbf{x}_m^{(n)} - \boldsymbol{\mu}_{\mathbf{x},i} \right) \\
&\qquad\qquad + Q \ln(2\pi) + \ln\left(\det\{\boldsymbol{\Sigma}_{\mathbf{x},i}\} \right) \Bigg) + \ln\{\psi_i\} \Bigg] \\
&\qquad + \sum_{m=L_{\mathrm{AR}}+1}^{M_n} \sum_{i=1}^{I} \Bigg[\eta_m^{(n,l)}(i) \\
&\qquad\qquad \cdot \left(-\frac{1}{2} \right) \Bigg(\left(\mathbf{x}_m^{(n)} - \sum_{\nu=1}^{L_{\mathrm{AR}}} \mathbf{A}_{i,\nu} \mathbf{x}_{m-\nu}^{(n)} - \mathbf{b}_i \right)^T \mathbf{V}_i^{-1} \left(\mathbf{x}_m^{(n)} - \sum_{\nu=1}^{L_{\mathrm{AR}}} \mathbf{A}_{i,\nu} \mathbf{x}_{m-\nu}^{(n)} - \mathbf{b}_i \right) \\
&\qquad\qquad + Q \ln(2\pi) + \ln\left(\det\{\mathbf{V}_i\} \right) \Bigg) + \sum_{k=1}^{I} \xi_m^{(n,l)}(k,i) \ln\{a_{k,i}\} \Bigg] \Bigg\}, \qquad \text{(A.12)}
\end{aligned}
$$

wobei $\det\{\cdot\}$ die Determinante einer Matrix bezeichnet. Die in diesem Ausdruck auftretenden bedingten Zustandswahrscheinlichkeiten $\eta_m^{(n,l)}(i)$ und $\xi_m^{(n,l)}(k,i)$ lassen sich sehr effizient durch eine modifizierte Version des BAUM-WELCH-Algorithmus [RJ93], welche im nächsten Unterabschnitt detailliert beschrieben wird, berechnen.

Berechnung der bedingten Zustandswahrscheinlichkeiten

Gemäß der Idee des BAUM-WELCH-Algorithmus [RJ93] werden die bedingten Zustandswahrscheinlichkeiten $\eta_m^{(n,l)}(i)$ und $\xi_m^{(n,l)}(k,i)$, die in (5.26) und (5.27) definiert sind, mit Hilfe der sogenannten Vorwärts- und Rückwärtswahrscheinlichkeiten

$$\alpha_m^{(n,l)}(i) := p\left(\mathbf{x}_{1:m}^{(n)}, \zeta_m^{(n)} = i \middle| \theta^{\{l\}} \right) \quad \text{für } 1 \le m \le M_n \qquad \text{(A.13)}$$

$$\beta_m^{(n,l)}(i) := p\left(\mathbf{x}_{m+1:M_n}^{(n)} \middle| \mathbf{x}_{m-L_{\mathrm{AR}}+1:m}^{(n)}, \zeta_m^{(n)} = i; \theta^{\{l\}} \right) \quad \text{für} \quad 1 \le m \le M_n \qquad \text{(A.14)}$$

gemäß

$$\eta_m^{(n,l)}(i) = \frac{p\left(\zeta_m^{(n)} = i, \mathbf{x}_{1:M_n}^{(n)} \middle| \theta^{\{l\}}\right)}{p\left(\mathbf{x}_{1:M_n}^{(n)}\right)} \tag{A.15}$$

$$\propto p\left(\mathbf{x}_{1:m}^{(n)}, \mathbf{x}_{m+1:M_n}^{(n)}, \zeta_m^{(n)} = i \middle| \theta^{\{l\}}\right) \tag{A.16}$$

$$\propto p\left(\mathbf{x}_{m+1:M_n}^{(n)} \middle| \mathbf{x}_{1:m}^{(n)}, \zeta_m^{(n)} = i; \theta^{\{l\}}\right) p\left(\mathbf{x}_{1:m}^{(n)}, \zeta_m^{(n)} = i \middle| \theta^{\{l\}}\right) \tag{A.17}$$

$$\propto p\left(\mathbf{x}_{m+1:M_n}^{(n)} \middle| \mathbf{x}_{m-L_{\mathrm{AR}}+1:m}^{(n)}, \zeta_m^{(n)} = i; \theta^{\{l\}}\right) p\left(\mathbf{x}_{1:m}^{(n)}, \zeta_m^{(n)} = i \middle| \theta^{\{l\}}\right) \tag{A.18}$$

$$\propto \beta_m^{(n,l)}(i)\, \alpha_m^{(n,l)}(i) \quad \text{für} \quad 1 \leq m \leq M_n \tag{A.19}$$

und

$$\xi_m^{(n,l)}(k,i) \tag{A.20}$$

$$= \frac{p\left(\zeta_m^{(n)} = i, \zeta_{m-1}^{(n)} = k, \mathbf{x}_{1:m}^{(n)}, \mathbf{x}_{m+1:M_n}^{(n)} \middle| \theta^{\{l\}}\right)}{p\left(\mathbf{x}_{1:M_n}^{(n)}\right)} \tag{A.21}$$

$$\propto p\left(\mathbf{x}_{m:M_n}^{(n)}, \zeta_m^{(n)} = i \middle| \zeta_{m-1}^{(n)} = k, \mathbf{x}_{1:m-1}^{(n)}; \theta^{\{l\}}\right) p\left(\zeta_{m-1}^{(n)} = k, \mathbf{x}_{1:m-1}^{(n)} \middle| \theta^{\{l\}}\right) \tag{A.22}$$

$$\propto p\left(\mathbf{x}_{m:M_n}^{(n)} \middle| \zeta_m^{(n)} = i, \zeta_{m-1}^{(n)} = k, \mathbf{x}_{1:m-1}^{(n)}; \theta^{\{l\}}\right) \cdot p\left(\zeta_m^{(n)} = i \middle| \zeta_{m-1}^{(n)} = k, \mathbf{x}_{1:m-1}^{(n)}; \theta^{\{l\}}\right) \alpha_{m-1}^{(n,l)}(k) \tag{A.23}$$

$$\propto p\left(\mathbf{x}_{m+1:M_n}^{(n)} \middle| \zeta_m^{(n)} = i, \zeta_{m-1}^{(n)} = k, \mathbf{x}_{1:m}^{(n)}; \theta^{\{l\}}\right) \cdot p\left(\mathbf{x}_m^{(n)} \middle| \zeta_m^{(n)} = i, \zeta_{m-1}^{(n)} = k, \mathbf{x}_{1:m-1}^{(n)}; \theta^{\{l\}}\right) a_{k,i}^{\{l\}} \alpha_{m-1}^{(n,l)}(k) \tag{A.24}$$

$$\propto p\left(\mathbf{x}_{m+1:M_n}^{(n)} \middle| \mathbf{x}_{m-L_{\mathrm{AR}}+1:m}^{(n)}, \zeta_m^{(n)} = i; \theta^{\{l\}}\right) \cdot p\left(\mathbf{x}_m^{(n)} \middle| \mathbf{x}_{m-L_{\mathrm{AR}}:m-1}^{(n)}, \zeta_m^{(n)} = i; \theta^{\{l\}}\right) a_{k,i}^{\{l\}} \alpha_{m-1}^{(n,l)}(k) \tag{A.25}$$

$$\propto \beta_m^{(n,l)}(i)\, p\left(\mathbf{x}_m^{(n)} \middle| \mathbf{x}_{m-L_{\mathrm{AR}}:m-1}^{(n)}, \zeta_m^{(n)} = i; \theta^{\{l\}}\right) a_{k,i}^{\{l\}} \alpha_{m-1}^{(n,l)}(k) \quad \text{für} \quad L_{\mathrm{AR}}+1 \leq m \leq M_n \tag{A.26}$$

ausgedrückt. Dabei ist zu berücksichtigen, dass hier und im weiteren Verlauf des Anhangs im Sinne einer besseren Lesbarkeit darauf verzichtet wurde, die Segmentindizes zur Kennzeichnung des zeitlichen Anfanges und Endes von Merkmalsvektorsequenzen derart zu beschränken, dass sie stets positiv sind. Im Falle von auftretenden nicht positiven Segmentindizes existieren die entsprechenden Merkmalsvektorsequenzen offensichtlich nicht und sind deshalb zu ignorieren.

Die zur eindeutigen Berechnung notwendigen Proportionalitätskonstanten lassen sich aus

den beiden Normierungsbedingungen

$$\sum_{i=1}^{I} \eta_m^{(n,l)}(i) = 1 \tag{A.27}$$

$$\sum_{k=1}^{I} \xi_m^{(n,l)}(k,i) = \eta_m^{(n,l)}(i) \tag{A.28}$$

bestimmen.

Der Vorteil der beiden Darstellungen (A.19) und (A.26) besteht nun darin, dass sich sowohl die Vorwärts- als auch Rückwärtswahrscheinlichkeiten rekursiv berechnen lassen. Dazu werden zunächst die Vorwärtswahrscheinlichkeiten für $1 \leq m \leq L_{\mathrm{AR}}$ und $i \in \{1,\ldots,I\}$ durch

$$\alpha_m^{(n,l)}(i) = p\left(\mathbf{x}_{1:m}^{(n)} \middle| \zeta_m^{(n)} = i; \theta^{\{l\}}\right) P\left(\zeta_m^{(n)} = i \middle| \theta^{\{l\}}\right) = \left[\prod_{m'=1}^{m} \mathcal{N}\left(\mathbf{x}_{m'}^{(n)}; \boldsymbol{\mu}_{\mathbf{x},i}^{\{l\}}, \boldsymbol{\Sigma}_{\mathbf{x},i}^{\{l\}}\right)\right] \psi_i^{\{l\}} \tag{A.29}$$

initialisiert. Anschließend wird ihre Berechnung für $m = L_{\mathrm{AR}}+1,\ldots,M_n$ und $i \in \{1,\ldots,I\}$ gemäß der Rekursion

$$\begin{aligned}
\alpha_m^{(n,l)}(i) &= p\left(\mathbf{x}_{1:m}^{(n)}, \zeta_m^{(n)} = i \middle| \theta^{\{l\}}\right) && \text{(A.30)}\\
&= \sum_{k=1}^{I} p\left(\mathbf{x}_{1:m}^{(n)}, \zeta_m^{(n)} = i, \zeta_{m-1}^{(n)} = k \middle| \theta^{\{l\}}\right) && \text{(A.31)}\\
&= \sum_{k=1}^{I} p\left(\mathbf{x}_m^{(n)} \middle| \zeta_m^{(n)} = i, \zeta_{m-1}^{(n)} = k, \mathbf{x}_{1:m-1}^{(n)}; \theta^{\{l\}}\right)\\
&\qquad \cdot P\left(\zeta_m^{(n)} = i \middle| \zeta_{m-1}^{(n)} = k, \mathbf{x}_{1:m-1}^{(n)}; \theta^{\{l\}}\right) P\left(\zeta_{m-1}^{(n)} = k, \mathbf{x}_{1:m-1}^{(n)} \middle| \theta^{\{l\}}\right) && \text{(A.32)}\\
&= \sum_{k=1}^{I} p\left(\mathbf{x}_m^{(n)} \middle| \mathbf{x}_{m-L_{\mathrm{AR}}:m-1}^{(n)}, \zeta_m^{(n)} = i; \theta^{\{l\}}\right) a_{k,i}^{\{l\}} \alpha_{m-1}^{(n,l)}(k) && \text{(A.33)}\\
&= \sum_{k=1}^{I} \mathcal{N}\left(\mathbf{x}_m^{(n)}; \sum_{\nu=1}^{L_{\mathrm{AR}}} \mathbf{A}_{i,\nu}^{\{l\}} \mathbf{x}_{m-\nu}^{(n)} + \mathbf{b}_i^{\{l\}}, \mathbf{V}_i^{\{l\}}\right) a_{k,i}^{\{l\}} \alpha_{m-1}^{(n,l)}(k) && \text{(A.34)}
\end{aligned}$$

durchgeführt, wobei für die letzte Umformung (5.13) verwendet wurde.

Die Initialisierung der Rückwärtswahrscheinlichkeiten für $i \in \{1,\ldots,I\}$ erfolgt durch

$$\beta_{M_n}^{(n,l)}(i) = 1. \tag{A.35}$$

Da sich die Rückwärtswahrscheinlichkeiten $\beta_m^{(n,l)}(i)$ für $m = M_n - 1, \ldots, 1$ gemäß

$$\beta_m^{(n,l)}(i) = p\left(\mathbf{x}_{m+1:M_n}^{(n)} \middle| \mathbf{x}_{m-L_{\mathrm{AR}}+1:m}^{(n)}, \zeta_m^{(n)} = i; \theta^{\{l\}}\right) \tag{A.36}$$

$$= \sum_{k=1}^{I} p\left(\mathbf{x}_{m+1:M_n}^{(n)} \middle| \mathbf{x}_{m-L_{\mathrm{AR}}+1:m}^{(n)}, \zeta_m^{(n)} = i, \zeta_{m+1}^{(n)} = k; \theta^{\{l\}}\right) \cdot P\left(\zeta_{m+1}^{(n)} = k \middle| \mathbf{x}_{m-L_{\mathrm{AR}}+1:m}^{(n)}, \zeta_m^{(n)} = i; \theta^{\{l\}}\right) \tag{A.37}$$

$$= \sum_{k=1}^{I} p\left(\mathbf{x}_{m+2:M_n}^{(n)} \middle| \mathbf{x}_{m-L_{\mathrm{AR}}+1:m+1}^{(n)}, \zeta_m^{(n)} = i, \zeta_{m+1}^{(n)} = k; \theta^{\{l\}}\right) \cdot p\left(\mathbf{x}_{m+1}^{(n)} \middle| \mathbf{x}_{m-L_{\mathrm{AR}}+1:m}^{(n)}, \zeta_m^{(n)} = i, \zeta_{m+1}^{(n)} = k; \theta^{\{l\}}\right) \cdot P\left(\zeta_{m+1}^{(n)} = k \middle| \mathbf{x}_{m-L_{\mathrm{AR}}+1:m}^{(n)}, \zeta_m^{(n)} = i; \theta^{\{l\}}\right) \tag{A.38}$$

$$= \sum_{k=1}^{I} p\left(\mathbf{x}_{m+2:M_n}^{(n)} \middle| \mathbf{x}_{m-L_{\mathrm{AR}}+2:m+1}^{(n)}, \zeta_{m+1}^{(n)} = k; \theta^{\{l\}}\right) \cdot p\left(\mathbf{x}_{m+1}^{(n)} \middle| \mathbf{x}_{m-L_{\mathrm{AR}}+1:m}^{(n)}, \zeta_{m+1}^{(n)} = k; \theta^{\{l\}}\right) \cdot P\left(\zeta_{m+1}^{(n)} = k \middle| \mathbf{x}_{m-L_{\mathrm{AR}}+1:m}^{(n)}, \zeta_m^{(n)} = i; \theta^{\{l\}}\right) \tag{A.39}$$

$$= \sum_{k=1}^{I} \beta_{m+1}^{(n,l)}(k)\, p\left(\mathbf{x}_{m+1}^{(n)} \middle| \mathbf{x}_{m-L_{\mathrm{AR}}+1:m}^{(n)}, \zeta_{m+1}^{(n)} = k; \theta^{\{l\}}\right) \cdot P\left(\zeta_{m+1}^{(n)} = k \middle| \mathbf{x}_{m-L_{\mathrm{AR}}+1:m}^{(n)}, \zeta_m^{(n)} = i; \theta^{\{l\}}\right) \tag{A.40}$$

audrücken lassen, ergibt sich daraus unter Verwendung des Modells (5.13) folgende Rekursionsvorschrift:

$$\beta_m^{(n,l)}(i) = \begin{cases} \sum\limits_{k=1}^{I} \beta_{m+1}^{(n,l)}(k)\, \mathcal{N}\left(\mathbf{x}_{m+1}^{(n)}; \sum\limits_{\nu=1}^{L_{\mathrm{AR}}} \mathbf{A}_{k,\nu}^{\{l\}} \mathbf{x}_{m+1-\nu}^{(n)} + \mathbf{b}_k^{\{l\}}, \mathbf{V}_k^{\{l\}}\right) a_{i,k}^{\{l\}} & \text{für } m \geq L_{\mathrm{AR}} \\ \sum\limits_{k=1}^{I} \beta_{m+1}^{(n,l)}(k)\, \mathcal{N}\left(\mathbf{x}_{m+1}^{(n)}; \boldsymbol{\mu}_{\mathbf{x},k}^{\{l\}}, \boldsymbol{\Sigma}_{\mathbf{x},k}^{\{l\}}\right) \psi_k & \text{für } m < L_{\mathrm{AR}} \end{cases}. \tag{A.41}$$

Aus der Definition der Vorwärtswahrscheinlichkeiten in (A.13) folgt weiterhin, dass sich die Likelihood für die n-te Merkmalsvektorsequenz mit ihrer Kenntnis gemäß

$$p\left(\mathbf{x}_{1:M_n}^{(n)} \middle| \theta^{\{l\}}\right) = \sum_{i=1}^{I} \alpha_{M_n}^{(n,l)}(i) \tag{A.42}$$

berechnen lässt. Weiterhin soll an dieser Stelle bemerkt werden, dass die Vorwärts- bzw. Rückwärtswahrscheinlichkeiten für wachsenden bzw. sinkende Segmentindizes approximativ exponentiell abnehmen und deshalb sehr kleine Werte annehmen können, so dass es sinnvoll ist, die Berechnung beider im logarithmischen Bereich durchzuführen.

A.1.2. *Maximization*-Schritt

Die Parametermenge $\theta^{\{l+1\}}$ wird nun im zweiten Schritt durch die Maximierung des Erwartungswertes der Loglikelihood gemäß

$$\theta^{\{l+1\}} = \underset{\theta}{\operatorname{argmax}}\, \mathscr{Q}_{l+1}(\theta) \tag{A.43}$$

bestimmt. Es kann gezeigt werden, dass die lokalen Maximumstellen von $\mathscr{Q}_{l+1}(\theta)$ gleichzeitig auch globale sind. Deshalb kann die Parametermenge $\theta^{\{l+1\}}$ durch die Suche der Nullstellen der partiellen Ableitungen von $\mathscr{Q}_{l+1}(\theta)$ nach den Komponenten von θ ermittelt werden.

Für diesen Zweck werden folgende Ableitungsregeln herangezogen, die für Vektoren $\mathbf{a}, \mathbf{b} \in \mathbb{R}^Q$ und Matrizen $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{Q\times Q}$ gelten, wobei $\mathbf{A}$ als symmetrisch und positiv definit vorausgesetzt wird [PP08, (51), (64), (78), (80)]:

$$\frac{\partial \ln(\det\{\mathbf{A}\})}{\partial \mathbf{A}} = \mathbf{A}^{-1} \tag{A.44}$$

$$\frac{\partial \mathbf{a}^T\mathbf{A}\mathbf{a}}{\partial \mathbf{A}} = \mathbf{a}\mathbf{a}^T \tag{A.45}$$

$$\frac{\partial (\mathbf{a}-\mathbf{b})^T \mathbf{A}(\mathbf{a}-\mathbf{b})}{\partial \mathbf{b}^T} = -2\mathbf{A}(\mathbf{a}-\mathbf{b}) \tag{A.46}$$

$$\frac{\partial (\mathbf{a}-\mathbf{B}\mathbf{b})^T \mathbf{A}(\mathbf{a}-\mathbf{B}\mathbf{b})}{\partial \mathbf{B}} = -2\mathbf{A}(\mathbf{a}-\mathbf{B}\mathbf{b})\mathbf{b}^T. \tag{A.47}$$

Bildet man die partielle Ableitung von $\mathscr{Q}_{l+1}(\theta)$ nach $\boldsymbol{\mu}_{\mathbf{x},i}^T$ unter Verwendung von (A.46), so erhält man

$$\frac{\partial \mathscr{Q}_{l+1}(\theta)}{\partial \boldsymbol{\mu}_{\mathbf{x},i}^T} = \boldsymbol{\Sigma}_{\mathbf{x},i}^{-1} \sum_{n=1}^{N}\sum_{m=1}^{L_{\mathrm{AR}}} \eta_m^{(n,l)}(i)\left(\mathbf{x}_m^{(n)} - \boldsymbol{\mu}_{\mathbf{x},i}^{\{l\}}\right). \tag{A.48}$$

Aus der Bedingung $\left.\frac{\partial \mathscr{Q}_{l+1}(\theta)}{\partial \boldsymbol{\mu}_{\mathbf{x},i}^T}\right|_{\boldsymbol{\mu}_{\mathbf{x},i}=\boldsymbol{\mu}_{\mathbf{x},i}^{\{l+1\}}} = \mathbf{0}$ folgt:

$$\boldsymbol{\mu}_{\mathbf{x},i}^{\{l+1\}} = \frac{\sum_{n=1}^{N}\sum_{m=1}^{L_{\mathrm{AR}}} \eta_m^{(n,l)}(i)\,\mathbf{x}_m^{(n)}}{\sum_{n=1}^{N}\sum_{m=1}^{L_{\mathrm{AR}}} \eta_m^{(n,l)}(i)}. \tag{A.49}$$

Die partielle Ableitungen von $\mathscr{Q}_{l+1}(\theta)$ nach $\boldsymbol{\Sigma}_{\mathbf{x},i}^{-1}$ und $\mathbf{V}_i^{-1}$ ergeben sich mit Berücksichtigung von (A.44), (A.45), der Tatsache, dass beide Matrizen $\boldsymbol{\Sigma}_{\mathbf{x},i}$ und $\mathbf{V}_i$ symmetrisch positiv definit sind, und

$$\ln(\det\{\mathbf{A}\}) = -\ln\left(\det\left\{\mathbf{A}^{-1}\right\}\right) \tag{A.50}$$

zu

$$\frac{\partial \mathscr{Q}_{l+1}(\theta)}{\partial \boldsymbol{\Sigma}_{\mathbf{x},i}^{-1}} = -\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{L_{\mathrm{AR}}} \eta_m^{(n,l)}(i)\left[\left(\mathbf{x}_m^{(n)} - \boldsymbol{\mu}_{\mathbf{x},i}^{\{l\}}\right)\left(\mathbf{x}_m^{(n)} - \boldsymbol{\mu}_{\mathbf{x},i}^{\{l\}}\right)^T - \boldsymbol{\Sigma}_{\mathbf{x},i}\right] \tag{A.51}$$

$$\frac{\partial \mathscr{Q}_{l+1}(\theta)}{\partial \mathbf{V}_i^{-1}} = -\frac{1}{2}\sum_{n=1}^{N}\sum_{m=L_{\mathrm{AR}}+1}^{M_n} \eta_m^{(n,l)}(i) \cdot \left[\left(\mathbf{x}_m^{(n)} - \sum_{\nu=1}^{L_{\mathrm{AR}}} \mathbf{A}_{i,\nu}^{\{l\}}\mathbf{x}_{m-\nu}^{(n)} - \mathbf{b}_i^{\{l\}}\right)\left(\mathbf{x}_m^{(n)} - \sum_{\nu=1}^{L_{\mathrm{AR}}} \mathbf{A}_{i,\nu}^{\{l\}}\mathbf{x}_{m-\nu}^{(n)} - \mathbf{b}_i^{\{l\}}\right)^T - \mathbf{V}_i\right]. \tag{A.52}$$

Die beiden Bedingungen $\left.\frac{\partial \mathscr{Q}_{l+1}(\theta)}{\partial \boldsymbol{\Sigma}_{\mathbf{x},i}^{-1}}\right|_{\boldsymbol{\Sigma}_{\mathbf{x},i}=\boldsymbol{\Sigma}_{\mathbf{x},i}^{\{l+1\}}} = \mathbf{0}$ und $\left.\frac{\partial \mathscr{Q}_{l+1}(\theta)}{\partial \mathbf{V}_i^{-1}}\right|_{\mathbf{V}_i=\mathbf{V}_i^{\{l+1\}}} = \mathbf{0}$ liefern

$$\boldsymbol{\Sigma}_{\mathbf{x},i}^{\{l+1\}} = \frac{\sum_{n=1}^{N}\sum_{m=1}^{L_{\mathrm{AR}}} \eta_m^{(n,l)}(i)\left(\mathbf{x}_m^{(n)} - \boldsymbol{\mu}_{\mathbf{x},i}^{\{l\}}\right)\left(\mathbf{x}_m^{(n)} - \boldsymbol{\mu}_{\mathbf{x},i}^{\{l\}}\right)^T}{\sum_{n=1}^{N}\sum_{m=1}^{L_{\mathrm{AR}}} \eta_m^{(n,l)}(i)} \tag{A.53}$$

$$\mathbf{V}_i^{\{l+1\}} = \frac{\sum_{n=1}^{N}\sum_{m=L_{\mathrm{AR}}+1}^{M_n} \eta_m^{(n,l)}(i)\left(\mathbf{x}_m^{(n)} - \sum_{\nu=1}^{L_{\mathrm{AR}}} \mathbf{A}_{i,\nu}^{\{l\}}\mathbf{x}_{m-\nu}^{(n)} - \mathbf{b}_i^{\{l\}}\right)\left(\mathbf{x}_m^{(n)} - \sum_{\nu=1}^{L_{\mathrm{AR}}} \mathbf{A}_{i,\nu}^{\{l\}}\mathbf{x}_{m-\nu}^{(n)} - \mathbf{b}_i^{\{l\}}\right)^T}{\sum_{n=1}^{N}\sum_{m=L_{\mathrm{AR}}+1}^{M_n} \eta_m^{(n,l)}(i)}. \tag{A.54}$$

Schließlich sind die partiellen Ableitungen von $\mathscr{Q}_{l+1}(\theta)$ nach $\mathbf{b}_i^T$ und $\mathbf{A}_{i,o}$ durch

$$\frac{\partial \mathscr{Q}_{l+1}(\theta)}{\partial \mathbf{b}_i^T} = \mathbf{V}_i^{-1}\sum_{n=1}^{N}\sum_{m=L_{\mathrm{AR}}+1}^{M_n} \eta_m^{(n,l)}(i)\left(\mathbf{x}_m^{(n)} - \sum_{\nu=1}^{L_{\mathrm{AR}}} \mathbf{A}_{i,\nu}\mathbf{x}_{m-\nu}^{(n)} - \mathbf{b}_i\right) \tag{A.55}$$

$$\frac{\partial \mathscr{Q}_{l+1}(\theta)}{\partial \mathbf{A}_{i,o}} = \mathbf{V}_i^{-1}\sum_{n=1}^{N}\sum_{m=L_{\mathrm{AR}}+1}^{M_n} \eta_m^{(n,l)}(i)\left(\mathbf{x}_m^{(n)} - \sum_{\nu=1}^{L_{\mathrm{AR}}} \mathbf{A}_{i,\nu}\mathbf{x}_{m-\nu}^{(n)} - \mathbf{b}_i\right)\left(\mathbf{x}_{m-o}^{(n)}\right)^T \tag{A.56}$$

gegeben, was aus (A.46) und (A.47) folgt. Die Bedingungen $\left.\frac{\partial \mathscr{Q}_{l+1}(\theta)}{\partial \mathbf{b}_i^T}\right|_{\mathbf{b}_i=\mathbf{b}_i^{\{l+1\}}} = \mathbf{0}$ und $\left.\frac{\partial \mathscr{Q}_{l+1}(\theta)}{\partial \mathbf{A}_{i,o}}\right|_{\mathbf{A}_{i,o}=\mathbf{A}_{i,o}^{\{l+1\}}} = \mathbf{0}$ für $o \in \{1,\ldots,L_{\mathrm{AR}}\}$ führen zu einem linearen Gleichungssystem, welches mit den abkürzenden Bezeichnungen

$$\left\langle \mathbf{x}^{\{l\}}(i)\right\rangle_{m':m''}^{[\nu,o]} = \sum_{n=1}^{N}\sum_{m=m'}^{m''} \eta_m^{(n,l)}(i)\,\mathbf{x}_{m-\nu}^{(n)}\left(\mathbf{x}_{m-o}^{(n)}\right)^T \tag{A.57}$$

$$\left\langle \mathbf{x}^{\{l\}}(i)\right\rangle_{m':m''}^{[\nu]} = \sum_{n=1}^{N}\sum_{m=m'}^{m''} \eta_m^{(n,l)}(i)\,\mathbf{x}_{m-\nu}^{(n)} \tag{A.58}$$

sowie den Matrizen

$$\mathbf{G}_i^{\{l\}} = \begin{bmatrix} \left\langle \mathbf{x}^{\{l\}}(i) \right\rangle_{L_{\mathrm{AR}}+1:M_n}^{[1,1]} & \cdots & \left\langle \mathbf{x}^{\{l\}}(i) \right\rangle_{L_{\mathrm{AR}}+1:M_n}^{[1,L_{\mathrm{AR}}]} & \left\langle \mathbf{x}^{\{l\}}(i) \right\rangle_{L_{\mathrm{AR}}+1:M_n}^{[1]} \\ \vdots & \ddots & \vdots & \vdots \\ \left\langle \mathbf{x}^{\{l\}}(i) \right\rangle_{L_{\mathrm{AR}}+1:M_n}^{[L_{\mathrm{AR}},1]} & \cdots & \left\langle \mathbf{x}^{\{l\}}(i) \right\rangle_{L_{\mathrm{AR}}+1:M_n}^{[L_{\mathrm{AR}},L_{\mathrm{AR}}]} & \left\langle \mathbf{x}^{\{l\}}(i) \right\rangle_{L_{\mathrm{AR}}+1:M_n}^{[L_{\mathrm{AR}}]} \\ \left(\left\langle \mathbf{x}^{\{l\}}(i) \right\rangle_{L_{\mathrm{AR}}+1:M_n}^{[1]} \right)^T & \cdots & \left(\left\langle \mathbf{x}^{\{l\}}(i) \right\rangle_{L_{\mathrm{AR}}+1:M_n}^{[L_{\mathrm{AR}}]} \right)^T & \sum_{n=1}^{N} \sum_{m=L_{\mathrm{AR}}+1}^{M_n} 1 \end{bmatrix} \tag{A.59}$$

und

$$\mathbf{H}_i^{\{l\}} = \begin{bmatrix} \left\langle \mathbf{x}^{\{l\}}(i) \right\rangle_{L_{\mathrm{AR}}+1:M_n}^{[1,0]} \\ \vdots \\ \left\langle \mathbf{x}^{\{l\}}(i) \right\rangle_{L_{\mathrm{AR}}+1:M_n}^{[L_{\mathrm{AR}},0]} \\ \left(\left\langle \mathbf{x}^{\{l\}}(i) \right\rangle_{L_{\mathrm{AR}}+1:M_n}^{[0]} \right)^T \end{bmatrix} \tag{A.60}$$

wie folgt geschrieben werden kann:

$$\mathbf{G}_i^{\{l\}} \begin{bmatrix} \left(\mathbf{A}_{i,1}^{\{l+1\}} \right)^T \\ \vdots \\ \left(\mathbf{A}_{i,L_{\mathrm{AR}}}^{\{l+1\}} \right)^T \\ (\mathbf{b}_i)^T \end{bmatrix} = \mathbf{H}_i^{\{l\}}. \tag{A.61}$$

Dazu ist zu bemerken, dass für den Fall, dass der Rang von $\mathbf{G}_i^{\{l\}} \in \mathbb{R}^{(L_{\mathrm{AR}}Q+1)\times(L_{\mathrm{AR}}Q+1)}$ kleiner als $L_{\mathrm{AR}}Q+1$ ist, bekanntlich unendlich viele Lösungen von (5.33) existieren. Da die Lösungsmenge aber zusammenhängend ist und für jede Lösung die entsprechenden partiellen Ableitungen verschwinden, ist jede Lösung auch eine lokale Maximumstelle von $\mathscr{Q}_{l+1}(\theta)$. In der Praxis wird der Einfachheit halber oft die Lösung mit der geringsten euklidischen Norm verwendet.

Die Maximierung von $\mathscr{Q}_{l+1}(\theta)$ bezüglich der Parameter ψ_i und $a_{k,i}$ muss unter Einhaltung der beiden Nebenbedingungen

$$\sum_{k=1}^{I} \psi_k = 1 \tag{A.62}$$

$$\sum_{i'=1}^{I} a_{k,i'} = 1 \quad \text{für } k \in \{1,\ldots,I\} \tag{A.63}$$

erfolgen, welche jeweils über die LAGRANGE-Multiplikatoren λ_1 und λ_2 berücksichtigt wer-

den, so dass sich die beiden Bedingungen

$$\frac{\partial}{\partial \psi_i}\left[\mathscr{Q}_{l+1}(\theta)+\lambda_1\left(\sum_{k=1}^{I}\psi_k-1\right)\right]\Bigg|_{\psi_i=\psi_i^{\{l+1\}}}=\sum_{n=1}^{N}\sum_{m=1}^{L_{\mathrm{AR}}}\eta_m^{(n,l)}(i)\frac{1}{\psi_i^{\{l+1\}}}+\lambda_1=0 \tag{A.64}$$

$$\frac{\partial}{\partial a_{k,i}}\left[\mathscr{Q}_{l+1}(\theta)+\lambda_2\left(\sum_{i'=1}^{I}a_{k,i'}-1\right)\right]\Bigg|_{a_{k,i}=a_{k,i}^{\{l+1\}}}=\sum_{n=1}^{N}\sum_{m=L_{\mathrm{AR}}+1}^{M_n}\xi_m^{(n,l)}(k,i)\frac{1}{a_{k,i}^{\{l+1\}}}+\lambda_2=0 \tag{A.65}$$

ergeben. Löst man die Gleichungen nach den gesuchten Parameter auf, so erhält man

$$\psi_i^{\{l+1\}}=-\frac{\sum_{n=1}^{N}\sum_{m=1}^{L_{\mathrm{AR}}}\eta_m^{(n,l)}(i)}{\lambda_1} \tag{A.66}$$

$$a_{k,i}^{\{l+1\}}=-\frac{\sum_{n=1}^{N}\sum_{L_{\mathrm{AR}}+1}^{M_n}\xi_m^{(n,l)}(k,i)}{\lambda_2}. \tag{A.67}$$

Die unbekannten LAGRANGE-Multiplikatoren können mit Hilfe der Summation von (A.66) und (A.67) über i unter Ausnutzung von (A.62), (A.27), (A.63) und (A.28) gemäß

$$1=\sum_{i=1}^{I}\psi_i^{\{l+1\}}=-\frac{\sum_{n=1}^{N}\sum_{m=1}^{L_{\mathrm{AR}}}\sum_{i=1}^{I}\eta_m^{(n,l)}(i)}{\lambda_1}=-\frac{N\cdot L_{\mathrm{AR}}}{\lambda_1} \tag{A.68}$$

$$1=\sum_{i=1}^{I}a_{k,i}^{\{l+1\}}=-\frac{\sum_{n=1}^{N}\sum_{L_{\mathrm{AR}}+1}^{M_n}\sum_{i=1}^{I}\xi_m^{(n,l)}(k,i)}{\lambda_2}=-\frac{\sum_{n=1}^{N}\sum_{L_{\mathrm{AR}}+1}^{M_n}\eta_{m-1}^{(n,l)}(k)}{\lambda_2} \tag{A.69}$$

ermittelt werden. Setzt man die resultierenden Lösungen für die LAGRANGE-Multiplikatoren

$$\lambda_1=-N\cdot L_{\mathrm{AR}} \tag{A.70}$$

$$\lambda_2=-\sum_{n=1}^{N}\sum_{L_{\mathrm{AR}}+1}^{M_n}\eta_{m-1}^{(n,l)}(k) \tag{A.71}$$

in (A.66) und (A.67) ein, gelangt man zu den gesuchten Parametern:

$$\psi_i^{\{l+1\}}=\frac{\sum_{n=1}^{N}\sum_{m=1}^{L_{\mathrm{AR}}}\eta_m^{(n,l)}(i)}{N\cdot L_{\mathrm{AR}}} \tag{A.72}$$

$$a_{k,i}^{\{l+1\}}=\frac{\sum_{n=1}^{N}\sum_{m=L_{\mathrm{AR}}+1}^{M_n}\xi_m^{(n,l)}(k,i)}{\sum_{n=1}^{N}\sum_{m=L_{\mathrm{AR}}+1}^{M_n}\eta_{m-1}^{(n,l)}(k)}. \tag{A.73}$$

Damit sind alle Komponenten von $\theta^{\{l+1\}}$ bestimmt und die $(l+1)$-te Iteration des *EM*-Algorithmus ist abgeschlossen.

A.2. Herleitungen und Beweise zum Beobachtungsmodell

A.2.1. Eigenschaften und Berechnung des Synthesefensters

Möchte man ein Signal gemäß (5.88) aus seinem Kurzzeit-Spektrum berechnen, wird ein Synthesefenster $w_{\mathrm{S}}(l')$ benötigt, welches die sogenannte Vollständigkeitsbedingung (5.85) erfüllt. An dieser Stelle soll gezeigt werden, dass sich diese Vollständigkeitsbedingung zu (5.87) vereinfacht, falls das Synthesefenster den gleichen Träger wie das Analysefenster besitzt, d.h. falls (5.86) erfüllt ist.

Dazu wird zunächst (5.85) gemäß

$$\left(\sum_{k=0}^{K-1} e^{j\frac{2\pi}{K}k(l-p')}\right) \sum_{m=-\infty}^{\infty} w_{\mathrm{S}}(l-mB)w_{\mathrm{A}}(p'-mB) = \delta\left(l-p'\right) \quad \text{für} \quad l,p' \in \mathbb{Z} \tag{A.74}$$

umformuliert. Da das Analyse- und Synthesefenster den gleichen Träger besitzen, d.h. dass (2.1) und (5.86) erfüllt sind, folgt $\forall B \in \mathbb{Z}$

$$w_{\mathrm{S}}(l-mB)w_{\mathrm{A}}(p'-mB) = 0 \text{ für } \left|l-p'\right| \geq L_w. \tag{A.75}$$

Damit ist (A.74) für $|l-p'| \geq L_w$ ohnehin erfüllt, so dass nur noch

$$\left(\sum_{k=0}^{K-1} e^{j\frac{2\pi}{K}k(l-p')}\right) \sum_{m=-\infty}^{\infty} w_{\mathrm{S}}(l-mB)w_{\mathrm{A}}(p'-mB) = \delta\left(l-p'\right) \quad \text{für} \quad \left|l-p'\right| < L_w \tag{A.76}$$

zu erfüllen ist. Unter Beachtung der Summenorthogonalität

$$\frac{1}{K}\sum_{k=0}^{K-1} e^{j\frac{2\pi}{K}k\mu} = \sum_{\nu=-\infty}^{\infty} \delta(\mu-\nu K) \quad \text{für } \mu \in \mathbb{Z}, K \in \mathbb{N} \tag{A.77}$$

und der Bedingung $L_w \leq K$ folgt, dass (A.76) auch für $l \neq p'$ erfüllt ist. Daher verbleibt nur noch die Bedingung

$$\sum_{m=-\infty}^{\infty} w_{\mathrm{S}}(l-mB)w_{\mathrm{A}}(l-mB) = \frac{1}{K} \quad \text{für } l \in \mathbb{Z}. \tag{A.78}$$

Da der linke Ausdruck in (A.78) die Periode B bezüglich l besitzt, genügt es, dass (A.78) nur für alle l innerhalb einer Periode erfüllt wird, so dass schließlich das zu zeigende Ergebnis

$$\sum_{m=-\infty}^{\infty} w_{\mathrm{S}}(l-mB)w_{\mathrm{A}}(l-mB) = \frac{1}{K} \quad \text{für } 0 \leq l < B \tag{A.79}$$

resultiert. Bedingt durch den Träger des Analyse- und Synthesefensters (siehe (2.1) und (5.86)) beinhaltet die Summe im linken Ausdruck von (A.79) nur endlich viele Summanden ungleich Null. Daher kann (A.79) auch äquivalent durch

$$\sum_{m=0}^{\alpha} w_{\mathrm{S}}(l-mB)w_{\mathrm{A}}(l-mB) = \frac{1}{K} \quad \text{für } 0 \leq l < B \tag{A.80}$$

mit

$$\alpha := \left\lfloor \frac{L_w}{B} \right\rfloor \tag{A.81}$$

ausgedrückt werden. An der Symmetrie dieser Bedingung bezüglich der Fenster erkennt man, dass die Bestimmung eines Synthesefensters zu einem gegebenen Analysefenster völlig analog zur Bestimmung eines Analysefensters zu einem gegebenen Synthesefenster verläuft.

Übrigens lässt sich die Bedingung (A.79) auch dadurch herleiten, dass (5.88) unter Anwendung der inversen diskreten FOURIER-Transformation (engl. ***I**nverse **D**iscrete* **F**OURIER ***T**ransform* (*IDFT*)) und durch Anwendung des Verschiebungssatzes gemäß

$$\frac{1}{K}\sum_{k=0}^{K-1} Y(m,k)\cdot e^{j\frac{2\pi}{K}k(l-mB)} = y_{w_\mathrm{A}}(m,l-mB), \tag{A.82}$$

wie folgt umformuliert wird

$$y(l) = \sum_{m=-\infty}^{\infty} w_\mathrm{S}(l-mB)Ky_{w_\mathrm{A}}(m,l-mB) \tag{A.83}$$

$$= y(l)\sum_{m=-\infty}^{\infty} w_\mathrm{S}(l-mB)Kw_\mathrm{A}(l-mB). \tag{A.84}$$

Zur Berechnung eines Synthesefensters lässt sich die Bedingung (A.80) in Matrixschreibweise wie folgt ausdrücken

$$\mathbf{W}_\mathrm{A}\mathbf{w}_\mathrm{S} = \frac{1}{K}\mathbf{1}, \tag{A.85}$$

wobei

$$\mathbf{1} := (1,...,1)^T \in \mathbb{R}^B \tag{A.86}$$

$$\mathbf{w}_\mathrm{S} := (w_\mathrm{S}(0),\ldots,w_\mathrm{S}(L_w-1))^T \in \mathbb{R}^{L_w} \tag{A.87}$$

$$\mathbf{W}_\mathrm{A} := \left(\mathbf{W}_\mathrm{A}^{(1)},...,\mathbf{W}_\mathrm{A}^{(\alpha+1)}\right) \in \mathbb{R}^{B\times L_w} \tag{A.88}$$

mit

$$\mathbf{W}_\mathrm{A}^{(i)} := \begin{cases} \mathrm{diag}\left\{w_\mathrm{A}((i-1)B),\ldots,w_\mathrm{A}(iB-1)\right\} \in \mathbb{R}^{B\times B} & \text{für } 1\le i\le\alpha \\ \begin{bmatrix} \mathrm{diag}\left\{w_\mathrm{A}(\alpha B),\ldots,w_\mathrm{A}(L_w-1)\right\} \\ \mathbf{0} \end{bmatrix} \in \mathbb{R}^{B\times(L_w-\alpha B)} & \text{für } i=\alpha+1 \end{cases}. \tag{A.89}$$

An dieser Stelle wird erkennbar, dass das Synthesefenster im Allgemeinen nicht eindeutig ist, da das Gleichungssystem unterbestimmt ist. Unter der Annahme, dass $\mathbf{W}_\mathrm{A}\mathbf{W}_\mathrm{A}^T$ nicht singulär ist, lässt sich jedoch die Lösung mit kleinster ℓ^2-Norm durch

$$\mathbf{w}_{\mathrm{S},\ell^2} = \frac{1}{K}\mathbf{W}_\mathrm{A}^T\left(\mathbf{W}_\mathrm{A}\mathbf{W}_\mathrm{A}^T\right)^{-1}\mathbf{1} \tag{A.90}$$

bestimmen. Unter Berücksichtigung von

$$\left(\mathbf{W}_{\mathrm{A}}\mathbf{W}_{\mathrm{A}}^{T}\right)^{-1} = \operatorname{diag}\left\{\frac{1}{\sum_{m=0}^{\alpha} w_{\mathrm{A}}^{2}(mB)},\ldots,\frac{1}{\sum_{m=0}^{\alpha} w_{\mathrm{A}}^{2}(B-1+mB)}\right\} \in \mathbb{R}^{B\times B}, \tag{A.91}$$

lässt sich (A.90) äquivalent durch

$$\begin{aligned}\mathbf{w}_{\mathrm{S},\ell^2} =& \frac{1}{K}\left(\frac{w_{\mathrm{A}}(0)}{\sum_{m=0}^{\alpha} w_{\mathrm{A}}^{2}(mB)},\ldots,\frac{w_{\mathrm{A}}(B-1)}{\sum_{m=0}^{\alpha} w_{\mathrm{A}}^{2}(B-1+mB)},\ldots,\right.\\ &\frac{w_{\mathrm{A}}((\alpha-1)B)}{\sum_{m=0}^{\alpha} w_{\mathrm{A}}^{2}(mB)},\ldots,\frac{w_{\mathrm{A}}(\alpha B-1)}{\sum_{m=0}^{\alpha} w_{\mathrm{A}}^{2}(B-1+mB)},\\ &\left.\frac{w_{\mathrm{A}}(\alpha B)}{\sum_{m=0}^{\alpha} w_{\mathrm{A}}^{2}(mB)},\ldots,\frac{w_{\mathrm{A}}(L_w-1)}{\sum_{m=0}^{\alpha} w_{\mathrm{A}}^{2}(L_w-1-\alpha B+mB)}\right)^{T}.\end{aligned} \tag{A.92}$$

ausdrücken.

A.2.2. Stauchungssatz für die zeitdiskrete FOURIER-Transformation

Satz 1. *Sei $x(l)$ ein zeitdiskretes Signal, welches die zeitdiskrete* FOURIER*-Transformation $X\left(e^{j\theta}\right)$ besitzt. Betrachtet werde nun ein weiteres zeitdiskretes Signal $y(l) := x(lB)$, welches durch Abtastung von $x(l)$ mit der Abtastfrequenz $\frac{1}{B}$, $B \in \mathbb{N}$, entsteht. Dessen zeitdiskrete* FOURIER*-Transformation $Y\left(e^{j\theta}\right)$ hängt dabei mit $X\left(e^{j\theta}\right)$ wie folgt zusammen:*

$$Y\left(e^{j\theta}\right) = \frac{1}{B}\sum_{m=0}^{B-1} X\left(e^{j\frac{1}{B}(\theta-2\pi m)}\right). \tag{A.93}$$

Beweis. Die inverse zeitdiskrete FOURIER-Transformation von $Y\left(e^{j\theta}\right)$ ist durch

$$y(l) = \frac{1}{2\pi}\int_{-\pi}^{\pi} Y\left(e^{j\theta}\right) e^{j\theta l} d\theta \tag{A.94}$$

gegeben. Setzt man (A.93) in (A.94) ein, so folgt

$$y(l) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{1}{B}\sum_{m=0}^{B-1} X\left(e^{j\frac{1}{B}(\theta-2\pi m)}\right) e^{j\theta l} d\theta. \tag{A.95}$$

Unter Verwendung der Variablensubstitution $\phi := \frac{1}{B}(\theta - 2\pi m)$ erhält man

$$y(l) = \frac{1}{2\pi}\int_{\frac{1}{B}(-\pi-2\pi m)}^{\frac{1}{B}(\pi-2\pi m)} \left(\frac{1}{B}\sum_{m=0}^{B-1} X\left(e^{j\phi}\right) e^{j(B\phi+2\pi m)l}\right) B d\phi \tag{A.96}$$

$$= \frac{1}{2\pi}\sum_{m=0}^{B-1}\int_{\frac{1}{B}(-\pi-2\pi m)}^{\frac{1}{B}(\pi-2\pi m)} X\left(e^{j\phi}\right) e^{j\phi B l} d\phi. \tag{A.97}$$

Beachtet man schließlich, dass die Grenzen der Integrale der Summanden jeweils aneinander stoßen, gelangt man zum gesuchten Ergebnis

$$y(l) = \frac{1}{2\pi} \int_{-\pi}^{\pi} X\left(e^{j\phi}\right) e^{j\phi Bl} d\phi \tag{A.98}$$

$$= x(lB). \tag{A.99}$$

□

A.2.3. Zusammenhang zwischen der Abklingkonstanten und der Nachhallzeit

In diesem Abschnitt wird ein Zusammenhang zwischen der mittleren Nachhallzeit T_{60} und der Abklingkonstanten τ_h hergeleitet, falls die Raumimpulsantwort einen Zufallsprozess darstellt, welcher durch (5.141) gegeben ist.

Die Nachhallzeit T_{60} ist als diejenige Zeit definiert, die benötigt wird, damit die Energie der Raumimpulsantwort um 60 dB abklingt. Nimmt man zur Vereinfachung bei der Behandlung von zeitdiskreten Signalen an, dass die Nachhallzeit T_{60} ein Vielfaches der Abtastdauer T_A darstellt, d.h. $T_{60} = l_0 T_\mathrm{A}$ mit $l_0 \in \mathbb{N}$, so muss l_0 die Bedingung

$$10\log_{10}\left(\frac{\mathrm{E}\left[\sum\limits_{l'=l}^{\infty} h^2(l')\right]}{\mathrm{E}\left[\sum\limits_{l'=0}^{\infty} h^2(l')\right]}\right) = -60 \tag{A.100}$$

erfüllen. In Anbetracht der Tatsache, dass es sich bei der Raumimpulsantwort nach Modell (5.141) um einen Zufallsprozess handelt, werden in (A.100) die Erwartungswerte der Energien verwendet.

Unter Verwendung von (5.146) und der Annahme, dass der Erwartungswert und der Limes vertauscht werden dürfen, erhält man

$$\mathrm{E}\left[\sum_{l'=l}^{\infty} \breve{h}^2(l')\right] = \mathrm{E}\left[\sum_{l'=0}^{\infty} \breve{h}^2(l')\right] - \mathrm{E}\left[\sum_{l'=0}^{l-1} \breve{h}^2(l')\right] \tag{A.101}$$

$$= \lim_{L_h\to\infty} \mathrm{E}\left[\sum_{l'=0}^{L_h-1} \breve{h}^2(l')\right] - \mathrm{E}\left[\sum_{l'=0}^{l-1} \breve{h}^2(l')\right] \tag{A.102}$$

$$= \lim_{L_h\to\infty} \sigma_h^2 \cdot \frac{e^{-\frac{2L_h}{\tau_h}} - 1}{e^{-\frac{2}{\tau_h}} - 1} - \sigma_h^2 \cdot \frac{e^{-\frac{2l}{\tau_h}} - 1}{e^{-\frac{2}{\tau_h}} - 1} \tag{A.103}$$

$$= \sigma_h^2 \cdot \frac{e^{-\frac{2l}{\tau_h}}}{1 - e^{-\frac{2}{\tau_h}}}. \tag{A.104}$$

Damit lässt sich der linke Term in (A.100) durch

$$10\log_{10}\left(\frac{\mathrm{E}\left[\sum\limits_{l'=l}^{\infty} h^2(l')\right]}{\mathrm{E}\left[\sum\limits_{l'=0}^{\infty} h^2(l')\right]}\right) = 10 \cdot \log_{10}\left(e^{-\frac{2l_0}{\tau_h}}\right) = 10 \cdot \frac{-\frac{2l_0}{\tau_h}}{\ln(10)} \tag{A.105}$$

ausdrücken, so dass nach dem Umstellen nach l_0 die Bedingung

$$l_0 = 3\tau_h \ln(10) \tag{A.106}$$

folgt. Nach einer Multiplikation beider Seite von (A.106) mit der Abtastdauer T_A ergibt sich der gesuchte Zusammenhang

$$T_{60} = l_0 T_\mathrm{A} = 3T_\mathrm{A}\tau_h \ln(10). \tag{A.107}$$

A.2.4. Herleitung der Erwartungswerte und Varianzen der Koeffizienten der Raumimpulsantwort im MEL-spektralen Bereich

In diesem Abschnitt werden die Erwartungswerte $\mu_{\breve{\bar{\mathscr{H}}}_{m',q}}$ sowie Varianzen $\sigma^2_{\breve{\bar{\mathscr{H}}}_{m',q}}$ der Koeffizienten der Raumimpulsantwort im MEL-spektralen Bereich $\bar{\mathscr{H}}_{m'}$ unter Annahme des vereinfachten Modells der Raumimpulsantwort (5.141) hergeleitet.

Für den Mittelwert ergibt sich zunächst gemäß der Definitionen (5.149) und (5.122)

$$\mu_{\breve{\bar{\mathscr{H}}}_{m',q}} = \mathrm{E}\left[\breve{\bar{\mathscr{H}}}_{m',q}\right] = \frac{1}{K_q^{(\mathrm{o})} - K_q^{(\mathrm{u})} + 1} \sum_{k=K_q^{(\mathrm{u})}}^{K_q^{(\mathrm{o})}} \mathrm{E}\left[\left|\breve{h}_{k,k}(m')\right|^2\right]. \tag{A.108}$$

Der Erwartungswert des Betragsquadrates der Band-zu-Band-Filter lässt sich mit Hilfe von (5.140) und des Modells der Raumimpulsantwort (5.141) gemäß

$$\mathrm{E}\left[\left|\breve{h}_{k,k}(m')\right|^2\right] = \mathrm{E}\left[\left|\sum_{p'=-L_w+1}^{L_w-1} w(p')\breve{h}(m'B+p')e^{-j\frac{2\pi}{K}kp'}\right|^2\right] \tag{A.109}$$

$$= \mathrm{E}\left[\left|\sum_{p'=-L_w+1}^{L_w-1} w(p')\cdot\sigma_h\cdot\breve{\nu}_h(m'B+p')\cdot\chi_h(m'B+p')\cdot e^{-\frac{m'B+p'}{\tau_h}} e^{-j\frac{2\pi}{K}kp'}\right|^2\right] \tag{A.110}$$

formulieren. Unter Verwendung der Abkürzungen

$$\delta_{m',p',k} := \sigma_h\cdot\chi_h(m'B+p')\cdot e^{-\frac{m'B+p'}{\tau_h}} w(p') e^{-j\frac{2\pi}{K}kp'} \tag{A.111}$$

$$\delta_{m',p'} := \left|\delta_{m',p',k}\right| = \sigma_h\cdot\chi_h(m'B+p')\cdot e^{-\frac{m'B+p'}{\tau_h}} w(p') \tag{A.112}$$

$$\breve{\nu}_{m',p'} := \breve{\nu}_h(m'B+p') \tag{A.113}$$

und der Korrelationsfunktion (5.142) des der Raumimpulsantwort zugrunde liegenden weißen, GAUSS'schen Zufallsprozesses $\breve{\nu}_h(l)$ lässt sich dieser Ausdruck zu

$$\mathrm{E}\left[\left|\breve{h}_{k,k}(m')\right|^2\right] = \mathrm{E}\left[\left|\sum_{p'=-L_w+1}^{L_w-1}\delta_{m',p',k}\breve{\nu}_{m',p'}\right|^2\right] = \sum_{p'=-L_w+1}^{L_w-1}\left|\delta_{m',p',k}\right|^2 = \sum_{p'=-L_w+1}^{L_w-1}\delta^2_{m',p'} \tag{A.114}$$

vereinfachen. Aufgrund der offensichtlichen Frequenzunabhängigkeit dieses Terms folgt für den Mittelwert $\mu_{\breve{\mathscr{H}}_{m',q}}$ mit (A.108)

$$\mu_{\breve{\mathscr{H}}_{m',q}} = \sum_{p'=-L_w+1}^{L_w-1} \delta^2_{m',p'}. \tag{A.115}$$

Für die Varianz $\sigma^2_{\breve{\mathscr{H}}_{m',q}}$ erhält man mit der Definition (5.150) und der Ausnutzung der Linearität des Erwartungswertes

$$\sigma^2_{\breve{\mathscr{H}}_{m',q}} = \mathrm{E}\left[\left(\breve{\mathscr{H}}_{m',q} - \mu_{\breve{\mathscr{H}}_{m',q}}\right)^2\right] = \mathrm{E}\left[\left(\breve{\mathscr{H}}_{m',q}\right)^2\right] - \left[\mu_{\breve{\mathscr{H}}_{m',q}}\right]^2, \tag{A.116}$$

wobei sich $\mathrm{E}\left[\left(\breve{\mathscr{H}}_{m',q}\right)^2\right]$ mit Hilfe von (5.122) gemäß

$$\mathrm{E}\left[\left(\breve{\mathscr{H}}_{m',q}\right)^2\right] = \left(\frac{1}{K_q^{(\mathrm{o})} - K_q^{(\mathrm{u})} + 1}\right)^2 \sum_{k=K_q^{(\mathrm{u})}}^{K_q^{(\mathrm{o})}} \sum_{k'=K_q^{(\mathrm{u})}}^{K_q^{(\mathrm{o})}} \mathrm{E}\left[\left|\breve{h}_{k,k}(m')\right|^2 \left|\breve{h}_{k',k'}(m')\right|^2\right] \tag{A.117}$$

ausdrücken lässt. Dabei lassen sich die einzelnen Summanden mit (5.140) und den Abkürzungen (A.111) und (A.113) durch

$$\begin{aligned}
&\mathrm{E}\left[\left|\breve{h}_{k,k}(m')\right|^2 \left|\breve{h}_{k',k'}(m')\right|^2\right] \\
&= \mathrm{E}\left[\left|\sum_{p'=-L_w+1}^{L_w-1} \delta_{m',p',k} \breve{\nu}_{m',p'}\right|^2 \left|\sum_{p''=-L_w+1}^{L_w-1} \delta_{m',p'',k'} \breve{\nu}_{m',p''}\right|^2\right] && \text{(A.118)} \\
&= \sum_{p',p'',p''',p''''=-L_w+1}^{L_w-1} \delta_{m',p',k} \delta^*_{m',p'',k} \delta_{m',p''',k'} \delta^*_{m',p'''',k'} \mathrm{E}\left[\breve{\nu}_{m',p'} \breve{\nu}_{m',p''} \breve{\nu}_{m',p'''} \breve{\nu}_{m',p''''}\right] && \text{(A.119)}
\end{aligned}$$

beschreiben. Unter Berücksichtigung der Tatsache, dass es sich bei $\breve{\nu}_h(l)$ um einen weißen, GAUSS'schen Zufallsprozess mit der Autokorrelationsfunktion (5.142) handelt, folgt mit der Definition (A.113) und [Iss18]

$$\mathrm{E}\left[\breve{\nu}_{m',p'} \breve{\nu}_{m',p''} \breve{\nu}_{m',p'''} \breve{\nu}_{m',p''''}\right] = \begin{cases} 3 & \text{für } p' = p'' = p''' = p'''' \\ 1 & \text{für } (p'' = p' \wedge p'''' = p''' \wedge p''' \neq p') \\ & \quad \vee (p''' = p' \wedge p'''' = p'' \wedge p'' \neq p') \\ & \quad \vee (p'''' = p' \wedge p''' = p'' \wedge p'' \neq p') \\ 0 & \text{sonst} \end{cases}. \tag{A.120}$$

Damit vereinfacht sich der Ausdruck (A.119) zu

$$\mathrm{E}\left[\left|\breve{h}_{k,k}(m')\right|^2 \left|\breve{h}_{k',k'}(m')\right|^2\right] = 3 \sum_{p'=-L_w+1}^{L_w-1} \delta^4_{m',p'} + \zeta^{(1)}_{m'} + \zeta^{(2)}_{m',k,k'}, \tag{A.121}$$

wobei

$$\zeta_{m'}^{(1)} := \sum_{p'=-L_w+1}^{L_w-1} \sum_{\substack{p'''=-L_w+1 \\ p'''\neq p'}}^{L_w-1} \delta_{m',p'}^2 \delta_{m',p'''}^2 \tag{A.122}$$

$$\begin{aligned}\zeta_{m',k,k'}^{(2)} := &\sum_{p'=-L_w+1}^{L_w-1} \sum_{\substack{p''=-L_w+1 \\ p''\neq p'}}^{L_w-1} \delta_{m',p',k} \delta_{m',p'',k}^* \delta_{m',p',k'} \delta_{m',p'',k'}^* \\ &+ \sum_{p'=-L_w+1}^{L_w-1} \sum_{\substack{p''=-L_w+1 \\ p''\neq p'}}^{L_w-1} \delta_{m',p',k} \delta_{m',p'',k}^* \delta_{m',p'',k'} \delta_{m',p',k'}^*. \end{aligned} \tag{A.123}$$

Stellt man mit Hilfe einfacher Umformungen $\zeta_{m'}^{(1)}$ und $\zeta_{m',k,k'}^{(2)}$ gemäß

$$\zeta_{m'}^{(1)} = \left(\sum_{p'=-L_w+1}^{L_w-1} \delta_{m',p'}^2 \right)^2 - \sum_{p'=-L_w+1}^{L_w-1} \delta_{m',p'}^4 \tag{A.124}$$

$$\zeta_{m',k,k'}^{(2)} = \left| \sum_{p'=-L_w+1}^{L_w-1} \delta_{m',p',k} \delta_{m',p',k'} \right|^2 + \left| \sum_{p'=-L_w+1}^{L_w-1} \delta_{m',p',k} \delta_{m',p',k'}^* \right|^2 - 2 \sum_{p'=-L_w+1}^{L_w-1} \delta_{m',p'}^4 \tag{A.125}$$

$$= \left| \sum_{p'=-L_w+1}^{L_w-1} \delta_{m',p',\frac{k+k'}{2}}^2 \right|^2 + \left| \sum_{p'=-L_w+1}^{L_w-1} \delta_{m',p',\frac{k-k'}{2}}^2 \right|^2 - 2 \sum_{p'=-L_w+1}^{L_w-1} \delta_{m',p'}^4 \tag{A.126}$$

dar und setzt das Resultat in (A.121) ein, dann erhält man

$$\begin{aligned}&\mathrm{E}\left[\left|\check{h}_{k,k}(m')\right|^2 \left|\check{h}_{k',k'}(m')\right|^2 \right] \\ &= \left(\sum_{p'=-L_w+1}^{L_w-1} \delta_{m',p'}^2 \right)^2 + \left| \sum_{p'=-L_w+1}^{L_w-1} \delta_{m',p',\frac{k+k'}{2}}^2 \right|^2 + \left| \sum_{p'=-L_w+1}^{L_w-1} \delta_{m',p',\frac{k-k'}{2}}^2 \right|^2 \end{aligned} \tag{A.127}$$

Durch das aufeinanderfolgende Einsetzen von (A.127) in (A.117) sowie (A.117) und (A.115) in (A.116) folgt der gesuchte vereinfachte Ausdruck für die Varianz

$$\sigma_{\check{\mathscr{H}}_{m',q}}^2 = \frac{1}{\left(K_q^{(\mathrm{o})} - K_q^{(\mathrm{u})} + 1\right)^2} \sum_{k,k'=K_q^{(\mathrm{u})}}^{K_q^{(\mathrm{o})}} \left(\left| \sum_{p'=-L_w+1}^{L_w-1} \delta_{m',p',\frac{k+k'}{2}}^2 \right|^2 + \left| \sum_{p'=-L_w+1}^{L_w-1} \delta_{m',p',\frac{k-k'}{2}}^2 \right|^2 \right). \tag{A.128}$$

A.2.5. Herleitung der Leistungskompensationskonstanten

Die Leistungskompensationskonstante C_{E} wird verwendet, um das Kurzzeit-Leistungsspektrum gemäß (5.117) zu approximieren. Sie soll dazu die Bedingung (5.119), welche äquiva-

lent gemäß

$$\mathrm{E}\left[\sum_{k',k''=0}^{K-1}\sum_{m',m''=-L_{H,\mathrm{u}}}^{L_H}\breve{X}(m-m',k')\breve{X}^*(m-m'',k'')\breve{h}_{k,k'}(m')\breve{h}^*_{k,k''}(m'')\right]$$

$$\stackrel{!}{=}\mathrm{E}\left[C_\mathrm{E}\cdot\sum_{m'=0}^{L_H}\left|\breve{X}(m-m',k)\right|^2\left|\breve{h}_{k,k}(m')\right|^2\right], \tag{A.129}$$

ausgedrückt werden kann, erfüllen. Dabei soll der Erwartungswert nicht nur über alle möglichen Eingangssignale gebildet werden, sondern ebenfalls über alle möglichen Impulsantworten, die sich gemäß dem vereinfachten Modell (5.141) ergeben können.

Um zu einer handhabbaren Lösung zu gelangen, wird in dieser Herleitung davon ausgegangen, dass es sich beim unverhallten Eingangssignal $\breve{x}(l)$ um einen reellen weißen GAUSS'schen Zufallsprozess handelt, welcher unkorreliert mit der Raumimpulsantwort ist und dessen Autokorrelationsfunktion

$$\mathrm{E}\left[\breve{x}(l)\breve{x}(l')\right]=\sigma_x^2\delta\left(l-l'\right) \tag{A.130}$$

erfüllt, wobei σ_x^2 die Leistung von $\breve{x}(l)$ bezeichnet. Die Autokorrelationsfunktion des Spektrums kann daher durch

$$\begin{aligned}&\mathrm{E}\left[\breve{X}(m-m',k')\breve{X}^*(m-m'',k'')\right]\\&=\sum_{l=0}^{L_w-1}\sum_{l'=0}^{L_w-1}w_\mathrm{A}(l)w_\mathrm{A}(l')\mathrm{E}\left[\breve{x}(l+\left(m-m'\right)B)\breve{x}(l'+\left(m-m''\right)B)\right]e^{-j\frac{2\pi}{K}(k'l-k''l')} &\text{(A.131)}\\&=\sigma_x^2\sum_{l=0}^{L_w-1}w_\mathrm{A}(l)w_\mathrm{A}(l+\left(m''-m'\right)B)e^{-j\frac{2\pi}{K}\{k'l-k''[l+(m''-m')B]\}} &\text{(A.132)}\end{aligned}$$

beschrieben werden. Weiterhin gilt für die Autokorrelationsfunktion der Raumimpulsantwort $\breve{h}(l)$ unter Berücksichtigung von (5.142)

$$\mathrm{E}\left[\breve{h}(l)\breve{h}(l')\right]=\sigma_h^2\delta\left(l-l'\right)\chi_h(l)e^{-\frac{2l}{\tau_h}}, \tag{A.133}$$

so dass sich die Autokorrelationsfunktion der Kreuzbandfilter mit Hilfe von (5.99) und (A.133) zu

$$\begin{aligned}&\mathrm{E}\left[\breve{h}_{k,k'}(m')\breve{h}^*_{k,k''}(m'')\right] &\text{(A.134)}\\&=\mathrm{E}\left[\sum_{l=0}^{L_h-1}\sum_{l'=0}^{L_h-1}\breve{h}(l)\breve{h}(l')\phi_{k,k'}(m'B-l)\phi^*_{k,k''}(m''B-l')\right] &\text{(A.135)}\\&=\sigma_h^2\sum_{l=0}^{L_h-1}\chi_h(l)\phi_{k,k'}(m'B-l)\phi^*_{k,k''}(m''B-l)e^{-\frac{2l}{\tau_h}}. &\text{(A.136)}\end{aligned}$$

ergibt. Beachtet man weiterhin, dass der Träger von $\phi_{k,k'}(l)$ durch $[-L_w+1,L_w-1]$ gegeben ist, so lässt sich (A.136) mit der Variablensubstitution $l'=m'B-l$ auch derart formulieren

$$\begin{aligned}&\mathrm{E}\left[\breve{h}_{k,k'}(m')\breve{h}^*_{k,k''}(m'')\right]\\&=\sigma_h^2\sum_{l'=-L_w+1}^{L_w-1}\phi_{k,k'}(l')\phi^*_{k,k''}(l'+\left(m''-m'\right)B)\cdot\chi_h(m'B-l')e^{-\frac{2\left(m'B-l'\right)}{\tau_h}}. &\text{(A.137)}\end{aligned}$$

Bildet man den Erwartungswert des Betragsquadrates von $Y(m,k)$, was dem linken Term in (A.129) entspricht, und setzt anschließend die gefundenen Ausdrücke (A.132) und (A.137) ein, so erhält man

$$\begin{aligned}
&\mathrm{E}\left[\left|\breve{Y}(m,k)\right|^2\right] \\
&= \sum_{m',m''=-L_{H,\mathrm{u}}}^{L_H} \sum_{k'=0}^{K-1}\sum_{k''=0}^{K-1} \mathrm{E}\left[\breve{X}(m-m',k')\breve{X}^*(m-m'',k'')\right] \mathrm{E}\left[\breve{h}_{k,k'}(m')\breve{h}^*_{k,k''}(m'')\right] \qquad \text{(A.138)} \\
&= \sigma_x^2\sigma_h^2 \sum_{m',m''=-L_{H,\mathrm{u}}}^{L_H} \sum_{l=0}^{L_w-1} w_{\mathrm{A}}(l) w_{\mathrm{A}}(l+\left(m''-m'\right)B) \sum_{l'=-L_w}^{L_w-1} \chi_h(m'B-l') e^{-\frac{2\left(m'B-l'\right)}{\tau_h}} \cdot \xi_{m''-m',l,l',k}, \qquad \text{(A.139)}
\end{aligned}$$

wobei

$$\xi_{m''-m',l,l',k} := \sum_{k'=0}^{K-1}\sum_{k''=0}^{K-1} \phi_{k,k'}(l')\phi^*_{k,k''}(l'+\left(m''-m'\right)B) e^{-j\frac{2\pi}{K}[k'l-k''(l+(m''-m')B)]}. \qquad \text{(A.140)}$$

Setzt man in (A.140) die Definition von $\phi_{k,k'}(l)$ gemäß (5.96) ein, so folgt

$$\begin{aligned}
\xi_{m''-m',l,l',k} &= \sum_{k'=0}^{K-1}\sum_{k''=0}^{K-1} \left[\sum_{p'=0}^{L_w-1} w_{\mathrm{A}}(p') w_{\mathrm{S}}(p'+l') e^{j\frac{2\pi}{K}k'(p'+l')} e^{-j\frac{2\pi}{K}kp'}\right] \\
&\quad \cdot \left[\sum_{p''=0}^{L_w-1} w_{\mathrm{A}}(p'') w_{\mathrm{S}}(p''+l'+\left(m''-m'\right)B) e^{-j\frac{2\pi}{K}k''(p''+l'+(m''-m')B)} e^{j\frac{2\pi}{K}kp''}\right] \\
&\quad \cdot e^{-j\frac{2\pi}{K}[k'l-k''(l+(m''-m')B)]} \qquad \text{(A.141)} \\
&= \sum_{p'=0}^{L_w-1} w_{\mathrm{A}}(p') w_{\mathrm{S}}(p'+l') e^{-j\frac{2\pi}{K}kp'} \sum_{p''=0}^{L_w-1} w_{\mathrm{A}}(p'') w_{\mathrm{S}}(p''+l'+\left(m''-m'\right)B) e^{j\frac{2\pi}{K}kp''} \\
&\quad \cdot \psi_{p',p'',l,l'} \qquad \text{(A.142)}
\end{aligned}$$

mit

$$\psi_{p',p'',l,l'} := \left[\sum_{k'=0}^{K-1} e^{j\frac{2\pi}{K}k'(p'+l'-l)}\right]\left[\sum_{k''=0}^{K-1} e^{-j\frac{2\pi}{K}k''(p''+l'-l)}\right] \qquad \text{(A.143)}$$

schreiben. Aufgrund der Summenorthogonalität (A.77) vereinfacht sich $\psi_{p',p'',l,l'}$ zu

$$\psi_{p',p'',l,l'} = K^2 \sum_{\nu'=-\infty}^{\infty}\sum_{\nu''=-\infty}^{\infty} \delta\left(p'+l'-l-\nu'K\right)\delta\left(p''+l'-l-\nu''K\right). \qquad \text{(A.144)}$$

Beachtet man noch die Identität

$$\delta(l-\mu)\,\delta\left(l-\mu'\right) = \delta(l-\mu)\,\delta\left(\mu-\mu'\right) \quad \text{für} \quad l,\mu,\mu' \in \mathbb{Z}, \qquad \text{(A.145)}$$

dann erhält man

$$\psi_{p',p'',l,l'} = K^2 \sum_{\nu'=-\infty}^{\infty}\sum_{\nu''=-\infty}^{\infty} \delta\left(p'+l'-l-\nu'K\right)\delta\left(p''-p'-\left(\nu''-\nu'\right)K\right). \qquad \text{(A.146)}$$

Da die Differenz $p'' - p'$ stets im Intervall $[-L_w+1, L_w-1]$ liegt und $K > L_w$ gilt, kann das Argument der zweiten DIRAC-Funktion in (A.146) überhaupt nur für $\nu'' = \nu'$ Null werden, so dass sich

$$\psi_{p',p'',l,l'} = K^2 \sum_{\nu'=-\infty}^{\infty} \delta\left(p' + l' - l - \nu' K\right) \delta\left(p'' - p'\right) \tag{A.147}$$

ergibt. Setzt man noch (A.147) in (A.142) ein, so erhält man den Ausdruck

$$\begin{aligned}
\xi_{m''-m',l,l',k} &= K^2 \sum_{\nu'=-\infty}^{\infty} \sum_{p'=0}^{L_w-1} w_{\mathrm{A}}(p') w_{\mathrm{S}}(p'+l') \sum_{p''=0}^{L_w-1} w_{\mathrm{A}}(p'') w_{\mathrm{S}}(p''+l'+\left(m''-m'\right)B) \\
&\quad \cdot \delta\left(p' + l' - l - \nu' K\right) \delta\left(p'' - p'\right) e^{-j\frac{2\pi}{K}k(p'-p'')} && \text{(A.148)} \\
&= K^2 \sum_{\nu'=-\infty}^{\infty} \sum_{p'=0}^{L_w-1} w_{\mathrm{A}}^2(p') w_{\mathrm{S}}(p'+l') \cdot w_{\mathrm{S}}(p'+l'+\left(m''-m'\right)B) \\
&\quad \cdot \delta\left(p' + l' - l - \nu' K\right) && \text{(A.149)} \\
&= K^2 \sum_{\nu'=-\infty}^{\infty} w_{\mathrm{A}}^2(-l'+l+\nu' K) w_{\mathrm{S}}(l+\nu' K) \cdot w_{\mathrm{S}}(l+\nu' K+\left(m''-m'\right)B), && \text{(A.150)}
\end{aligned}$$

woran zu erkennen ist, dass $\xi_{m''-m',l,l',k}$ gar nicht von k abhängt. Für $l \in [-L_w+1, L_w-1]$ gilt $w_{\mathrm{S}}(l+\nu' K) = 0 \; \forall \nu' \neq 0$, so dass

$$\xi_{m''-m',l,l',k} = K^2 w_{\mathrm{A}}^2(-l'+l) w_{\mathrm{S}}(l) w_{\mathrm{S}}(l+\left(m''-m'\right)B). \tag{A.151}$$

Setzt man (A.151) in (A.139) ein, so ergibt sich

$$\mathrm{E}\left[\left|\check{Y}(m,k)\right|^2\right] = \sigma_x^2 \sigma_h^2 \cdot C_{\mathrm{Z}} \tag{A.152}$$

mit

$$\begin{aligned}
C_{\mathrm{Z}} := K^2 &\sum_{m',m''=-L_{H,\mathrm{u}}}^{L_H} \sum_{l=0}^{L_w-1} w_{\mathrm{A}}(l) w_{\mathrm{S}}(l) w_{\mathrm{A}}(l+\left(m''-m'\right)B) w_{\mathrm{S}}(l+\left(m''-m'\right)B) \\
&\cdot \sum_{l'=-L_w+1}^{L_w-1} \chi_h(m'B-l') e^{-\frac{2(m'B-l')}{\tau_h}} w_{\mathrm{A}}^2(-l'+l). && \text{(A.153)}
\end{aligned}$$

Der rechte Ausdruck in (A.129) lässt sich mit (A.132) und (A.137) zu

$$\begin{aligned}
&\mathrm{E}\left[C_{\mathrm{E}} \cdot \sum_{m'=0}^{L_H} \left|\check{X}(m-m',k)\right|^2 \left|\check{h}_{k,k}(m')\right|^2\right] \\
&= C_{\mathrm{E}} \sum_{m'=0}^{L_H} \mathrm{E}\left[\left|\check{X}(m-m',k)\right|^2\right] \mathrm{E}\left[\left|\check{h}_{k,k}(m')\right|^2\right] && \text{(A.154)} \\
&= C_{\mathrm{E}} \sum_{m'=0}^{L_H} \left(\sigma_x^2 \sum_{l=0}^{L_w-1} w_{\mathrm{A}}^2(l)\right) \cdot \left(\sigma_h^2 \sum_{l'=-L_w+1}^{L_w-1} \left|\phi_{k,k}(l')\right|^2 \chi_h(m'B-l') e^{-\frac{2(m'B-l')}{\tau_h}}\right) && \text{(A.155)} \\
&= C_{\mathrm{E}} \cdot \sigma_x^2 \sigma_h^2 \cdot C_{\mathrm{N}} && \text{(A.156)}
\end{aligned}$$

vereinfachen, wobei C_{N} unter Beachtung von (5.96) durch

$$C_{\mathrm{N}} := \left(\sum_{l=0}^{L_w-1} w_{\mathrm{A}}^2(l)\right)\left[\sum_{m'=0}^{L_H}\sum_{l'=-L_w+1}^{L_w-1}\left(\sum_{p''=0}^{L_w-1} w_{\mathrm{A}}(p'')w_{\mathrm{S}}(p''+l')\right)^2 \chi_h(m'B-l')e^{-\frac{2(m'B-l')}{\tau_h}}\right] \tag{A.157}$$

definiert ist.

Die gesuchte Leistungskompensationskonstante C_{E} resultiert schließlich aus dem Gleichsetzen der beiden Ausdrücke (A.156) und (A.152):

$$C_{\mathrm{E}} = \frac{C_{\mathrm{Z}}}{C_{\mathrm{N}}}. \tag{A.158}$$

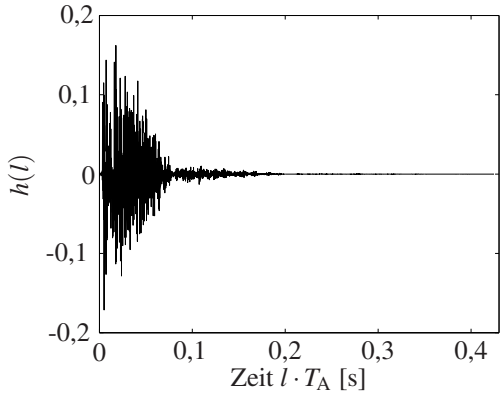
A.3. Raumimpulsantworten zur Erzeugung der AURORA5-Datenbank

Bei der ursprünglichen Erstellung der AURORA5-Datenbank [Hir07] wurden zwei unterschiedliche simulierte Freisprechumgebungen betrachtet, welche stellvertretend als Büro und Wohnzimmer bezeichnet wurden. Für jeden dieser zwei Räume wurden zunächst 3 unterschiedliche RIAs erzeugt, welche jeweils 3 unterschiedliche Ausprägungen bzw. Beschaffenheiten dieser Räume repräsentieren sollten. Dabei wiesen die 3 RIAs für das Büro Nachhallzeiten T_{60} von etwa 0,3 s, 0,35 s und 0,4 s und entsprechende *DRR*s von etwa −6,0 dB, −6,4 dB und −6,8 dB auf. Beim Wohnzimmer nahm die Nachhallzeit Werte von etwa 0,4 s, 0,45 s und 0,5 s an, wobei die entsprechenden *DRR*s etwa −5,7 dB, −6,5 dB und −7,0 dB betrugen. Zur Berechnung aller 6 RIAs wurde im Wesentlichen der direkte Anteil samt den frühen Reflexionen mit Hilfe der Spiegelquellenmethode [All79] erzeugt, wobei anschließend der Anteil des späten Nachhalls künstlich hinzugefügt wurde. Für weitere Details sei auf die ausführlichere Dokumentation in [HF05] verwiesen. Die verhallten Testdaten für jeden der zwei Räume wurden anschließend dadurch erzeugt, indem saubere Sprachsignale der *TI*-Digits-Datenbank mit jeweils einer der 3 raumspezifischen RIAs gefaltet wurden.

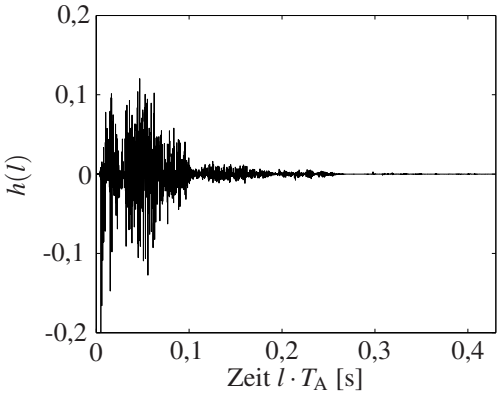
Die insgesamt 6 RIAs sind in Abb. A.1 illustriert. Zudem zeigt Abb. A.2 die entsprechenden log-MEL-spektrale Repräsentationen $\bar{h}_{m,q}$.

A.4. Statistische Signifikanz der Unterschiede zwischen Wortfehlerraten

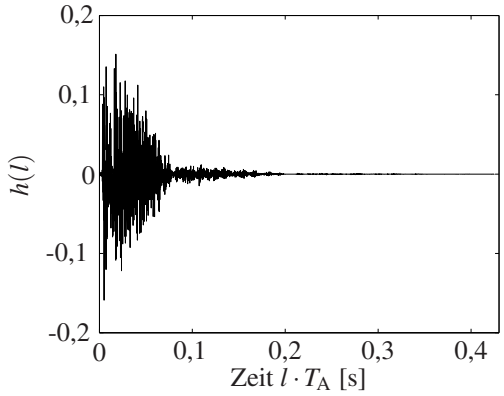
Zur approximativen Untersuchung der statistischen Signifikanz der Unterschiede der Wortfehlerraten zweier Verfahren, $\lambda_{w,1}$ und $\lambda_{w,2}$, sei hier nur eine stark vereinfachte Methode aus [GC89] angegeben, deren Defizite im Anschluss diskutiert werden sollen. Diese geht von der Annahme aus, dass es sich bei der Erkennungsaufgabe um ein BERNOULLI-Experiment bestehend aus N_{Ges} unabhängigen Einzelexperimenten handelt, bei dem jeweils ein Wort entweder falsch oder richtig erkannt werden kann. Die Wahrscheinlichkeit ein Wort richtig zu erkennen liegt bei den beiden Verfahren jeweils näherungsweise bei $\lambda_{w,1}$ bzw. $\lambda_{w,2}$. Bei diesen beiden Wahrscheinlichkeiten handelt es sich um Schätzungen, wobei sich die Varianzen



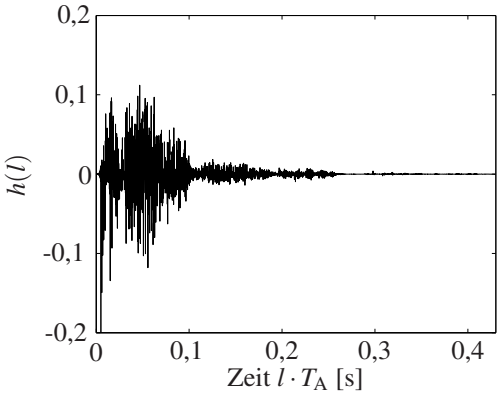
(a) Büro ($T_{60} \approx 0{,}3\,\mathrm{s}$, DRR $\approx -6{,}0\,\mathrm{dB}$)



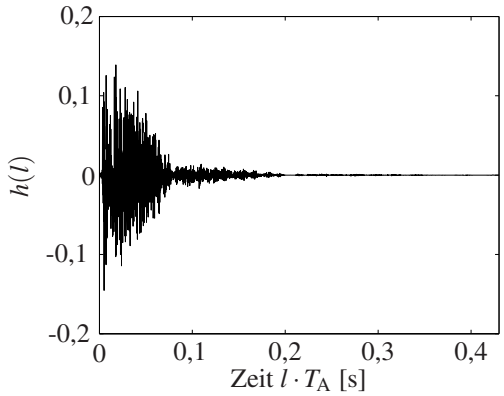
(b) Wohnzimmer ($T_{60} \approx 0{,}4\,\mathrm{s}$, DRR $\approx -5{,}7\,\mathrm{dB}$)



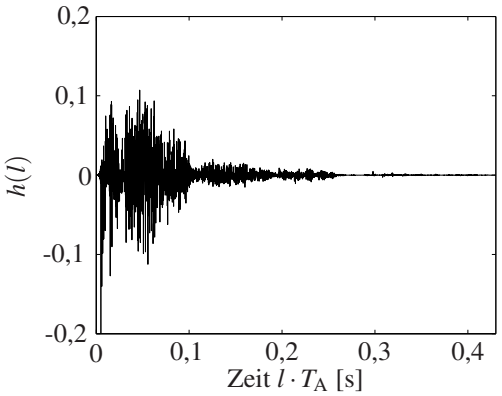
(c) Büro ($T_{60} \approx 0{,}35\,\mathrm{s}$, DRR $\approx -6{,}4\,\mathrm{dB}$)



(d) Wohnzimmer ($T_{60} \approx 0{,}45\,\mathrm{s}$, DRR $\approx -6{,}5\,\mathrm{dB}$)

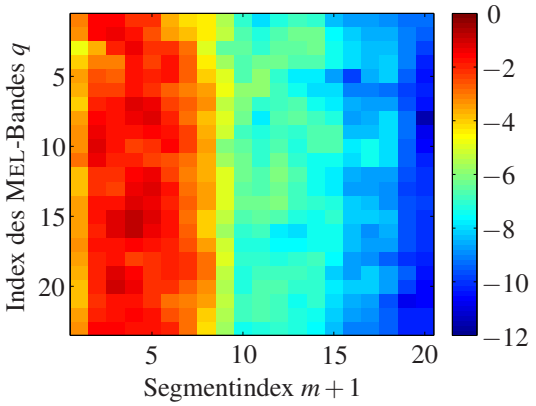


(e) Büro ($T_{60} \approx 0{,}4\,\mathrm{s}$, DRR $\approx -6{,}8\,\mathrm{dB}$)

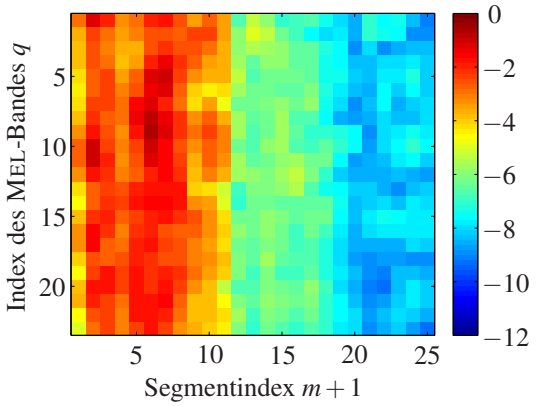


(f) Wohnzimmer ($T_{60} \approx 0{,}5\,\mathrm{s}$, DRR $\approx -7{,}0\,\mathrm{dB}$)

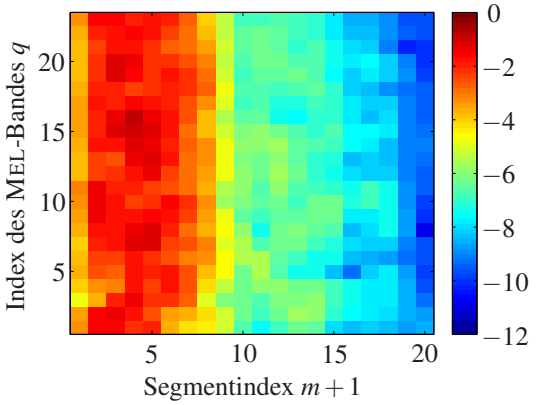
Abbildung A.1.: *Zur Erstellung der AURORA5-Datenbank verwendete RIAs.*



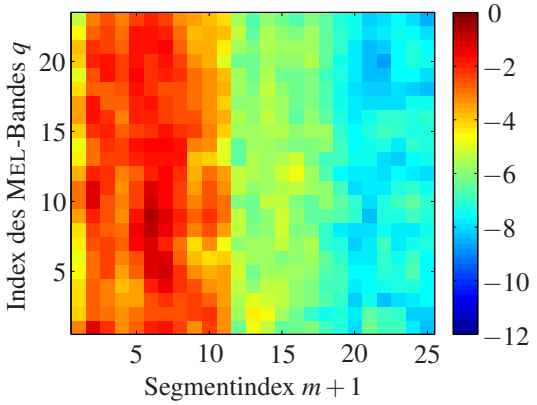
(a) Büro ($T_{60} \approx 0{,}3\,\mathrm{s}$, DRR $\approx -6{,}0\,\mathrm{dB}$)



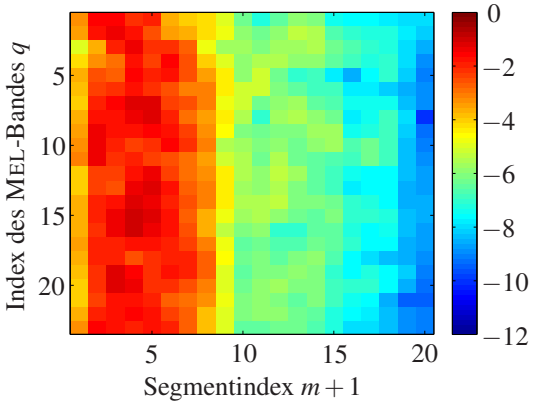
(b) Wohnzimmer ($T_{60} \approx 0{,}4\,\mathrm{s}$, DRR $\approx -5{,}7\,\mathrm{dB}$)



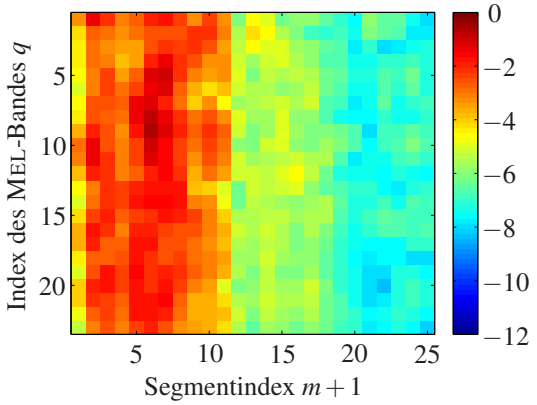
(c) Büro ($T_{60} \approx 0{,}35\,\mathrm{s}$, DRR $\approx -6{,}4\,\mathrm{dB}$)



(d) Wohnzimmer ($T_{60} \approx 0{,}45\,\mathrm{s}$, DRR $\approx -6{,}5\,\mathrm{dB}$)



(e) Büro ($T_{60} \approx 0{,}4\,\mathrm{s}$, DRR $\approx -6{,}8\,\mathrm{dB}$)



(f) Wohnzimmer ($T_{60} \approx 0{,}5\,\mathrm{s}$, DRR $\approx -7{,}0\,\mathrm{dB}$)

Abbildung A.2.: *Log-*MEL*-spektrale Repräsentationen $\bar{h}_{m,q}$ der RIAs, die ursprünglich zur Erstellung der AURORA5-Datenbank verwendet worden sind.*

des Schätzfehlers bedingt durch das BERNOULLI-Experiment gemäß

$$\sigma^2_{\lambda_{w,j}} = \frac{\lambda_{w,j}\left(1-\lambda_{w,j}\right)}{N_{\text{Ges}}} \quad \text{für } j = 1,2 \tag{A.159}$$

berechnen lassen. Aufgrund der sehr hohen Anzahl an Einzelexperimenten N_{Ges} können die Schätzfehler unter Beachtung des Zentralen Grenzwertsatzes [Man64] als annähernd normalverteilt angesehen werden. Unter der Nullhypothese, dass beide Verfahren im Mittel dieselbe Fehlerrate liefern, und der weiteren Annahme, dass die Schätzfehler beider Verfahren unabhängig sind, ist die Differenz $\Delta_{\lambda_w} := \lambda_{w,1} - \lambda_{w,2}$ ebenfalls normalverteilt mit der Varianz $\sigma^2_{\lambda_{w,1}} + \sigma^2_{\lambda_{w,2}}$. In diesem Fall ist der Unterschied zwischen den Wortfehlerraten der beiden betrachteten Verfahren dann als statistisch signifikant mit einem Signifikanzniveau von 5% anzusehen, wenn die Differenz Δ_{λ_w} außerhalb des 95%-Konfidenzintervalls

$$\mathscr{I}_{95\%} := \left[-\sqrt{\sigma^2_{\lambda_{w,1}} + \sigma^2_{\lambda_{w,2}}}, \sqrt{\sigma^2_{\lambda_{w,1}} + \sigma^2_{\lambda_{w,2}}}\right], \tag{A.160}$$

liegt.

Diese Art des Signifikanztests geht jedoch von Annahmen aus, die für die in dieser Arbeit betrachteten Testszenarien im Allgemeinen nicht zutreffend sind. So ist die Annahme, dass die Erkennungsergebnisse für einzelne Wörter als unabhängige Ereignisse angesehen werden können, wenn überhaupt nur für die Einzelworterkennung, die mit der AURORA5-Datenbank durchgeführt wird, gerechtfertigt. Für Erkennungsaufgaben, die im Zusammenhang mit der AURORA4-Datenbank stehen und bei denen ein Sprachmodell verwendet wird, besteht offensichtlich eine Abhängigkeit zwischen aufeinanderfolgenden erkannten Wörtern. Aber auch im Falle der Einzelworterkennung muss berücksichtigt werden, dass pro Wort mehrere Einfügefehler auftreten können, so dass die Wahrscheinlichkeiten für die richtige Erkennung eines Wortes in der Regel von Wort zu Wort variieren. Eine weitere unzutreffende Annahme ist die Unabhängigkeit der Schätzfehler beider Verfahren, da beiden Verfahren dieselben oder zumindest sehr ähnliche Testdaten zugrunde liegen. Daher kann davon ausgegangen werden, dass aufgrund der Ähnlichkeit beider Verfahren eine Ähnlichkeit der Fehler zu erwarten ist.

Eine Möglichkeit zur Berücksichtigung der Abhängigkeit der Ergebnisse zweier Verfahren bietet der sogenannte MCNEMAR'sche Test [GC89]. Dabei können Aussagen über die relative Leistungsfähigkeit zweier Verfahren beruhend auf der Information darüber gemacht werden, wie viele Wörter des Testdatensatzes existieren, die vom ersten Verfahren richtig und vom zweiten falsch erkannt wurden, und umgekehrt. Zur Lösung des Problems der Abhängigkeit von aufeinanderfolgenden Wörtern bietet sich der sogenannte Test mit gepaarten Stichproben (engl. *matched pairs test*) [GC89] an, bei dem die Testdaten in unabhängige Segmente wie einzelne Sätze unterteilt werden und anschließend die durchschnittliche Anzahl der Fehler pro Segment beider Verfahren verglichen wird. Auf die Durchführung von Signifikanztests dieser Art wurde in dieser Arbeit jedoch verzichtet, da das primäre Ziel in der Feststellung von groben Tendenzen lag und nicht in der Interpretation von marginalen, eventuell signfikanten, Unterschieden.

Formelzeichen

Allgemeine Bemerkungen

- Wahrscheinlichkeiten werden durchgehend durch $P(\cdot)$ gekennzeichnet, Verteilungsdichtefunktionen hingegen durch $p(\cdot)$. Dabei wird von der in der Literatur häufig verwendeten Notation, die Zufallsvariable als Index zu verwenden, zugunsten der Lesbarkeit der Ausdrücke in den Fällen abgesehen, wo die Zufallsvariable aus dem Argument der Verteilungsdichtefunktion ersichtlich wird.
- Für den Erwartungswert einer Zufallsvariable wird die Notation $\mathrm{E}[\cdot]$ verwendet. Um deutlich zu machen, bezüglich welcher Zufallsvariablen der Erwartungswert zu bilden ist, wird die entsprechende Zufallsvariable als Index verwendet.
- Zufallsvariablen werden stets mit einem Breve gemäß $\breve{(\cdot)}$ versehen. Die entsprechenden Realisierungen tragen dasselbe Symbol, jedoch ohne das Breve.
- Geschätzte Werte werden stets durch ein zusätzliches Dach gemäß $\hat{(\cdot)}$ gekennzeichnet.

Spezielle Symbole und Definitionen

$\mathbf{1}$ ..... Vektor bestehend aus Einsen
$\mathbf{I}$ ..... Einheitsmatrix
$\mathbf{0}$ ..... Nullvektor
$*$ ..... Faltung
$(\cdot)^*$ ..... Komplexe Konjugation
$(\cdot)^T$ ..... Transposition
$\lfloor \cdot \rfloor$ ..... Rundung auf die nächstkleinere oder gleich große, ganze Zahl
$\det\{\cdot\}$ ..... Determinante
$\Re[\cdot]$ ..... Realteil

- Zeitdiskreter δ-Impuls:

$$\delta(l) := \begin{cases} 1 & \text{für } l = 0 \\ 0 & \text{für } l \in \mathbb{Z}\backslash\{0\} \end{cases}. \tag{161}$$

- Diagonalmatrix oder Vektor bestehend aus den Diagonalelementen einer Matrix:

  Das Ergebnis der in dieser Arbeit verwendeten Operation $\mathrm{diag}\{\cdot\}$ hängt vom Typ ihres Argumentes ab. Handelt es sich beim Argument um einen Vektor, so ist das

Ergebnis eine Diagonalmatrix mit den Elementen des Vektors auf der Hauptdiagonalen gemäß

$$\operatorname{diag}\left\{(x_1,x_2,\ldots,x_{N-1},x_N)^T\right\} := \begin{bmatrix} x_1 & 0 & 0 & \ldots & 0 \\ 0 & x_2 & 0 & \ldots & 0 \\ 0 & 0 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & x_{N-1} & 0 \\ 0 & 0 & \ldots & 0 & x_N \end{bmatrix}. \tag{162}$$

Ist das Argument jedoch eine Matrix, so liefert die Anwendung von $\operatorname{diag}\{\cdot\}$ einen Vektor, dessen Einträge aus den Elementen der Hauptdiagonalen der Matrix bestehen:

$$\operatorname{diag}\left\{\begin{bmatrix} x_{1,1} & x_{1,2} & x_{1,3} & \ldots & x_{1,N} \\ x_{2,1} & x_{2,2} & x_{2,3} & \ldots & x_{2,N} \\ x_{3,1} & x_{3,2} & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & x_{N-1,N-1} & x_{N-1,N} \\ x_{N,1} & x_{N,2} & \ldots & x_{N,N-1} & x_{N,N} \end{bmatrix}\right\} := \begin{bmatrix} x_{1,1} \\ x_{2,2} \\ \vdots \\ x_{N-1,N-1} \\ x_{N,N} \end{bmatrix}. \tag{163}$$

- Blockdiagonalmatrix:

$$\operatorname{blockdiag}\{\mathbf{A}_1,\mathbf{A}_2,\ldots,\mathbf{A}_{N-1},\mathbf{A}_N\} := \begin{bmatrix} \mathbf{A}_1 & \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_2 & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \mathbf{A}_{N-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0} & \mathbf{A}_N \end{bmatrix}. \tag{164}$$

Römische Formelzeichen

$a_{i,k}$ ..... Wahrscheinlichkeit für den Wechsel von dem i-ten zum k-ten Teilmodell eines *SLDM*

$a_{i,k}^{\{l\}}$ ..... Wahrscheinlichkeit für den Wechsel von dem i-ten zum k-ten Teilmodell eines *SLDM* berechnet nach der l-ten *EM*-Iteration

$\mathbf{A}_{i,\nu}$ ..... Zustandsübergangsmatrix des i-ten Teilmodells eines *SLDM* für den Versatzindex ν

$\mathbf{A}_{i,\nu}^{\{l\}}$ ..... Zustandsübergangsmatrix des i-ten Teilmodells eines *SLDM* für den Versatzindex ν berechnet nach der l-ten *EM*-Iteration

$\mathbf{b}_i$ ..... Biaskorrekturvektor des i-ten Teilmodells eines *SLDM*

$\mathbf{b}_i^{\{l\}}$ ..... Biaskorrekturvektor des i-ten Teilmodells eines *SLDM* berechnet nach der l-ten *EM*-Iteration

B ..... Fenstervorschub (bei der Merkmalsextraktion)

C_{E} ..... Multiplikative Konstante zur Kompensation der Fehler bei der approximativen Darstellung des Kurzzeit-Leistungsspektrums eines verhallten Signals

C_{50} ..... Klarheitsmaß zur Beschreibung der Verständlichkeit von Sprache

C_{80} ..... Klarheitsmaß zur Beschreibung der Durchsichtigkeit von Musik

D_{INIT} ..... Zu minimierender Gesamtabstand bei der *GMM*-Initialisierung (definiert in (5.50))

DRR ..... Verhältnis zwischen der Energie des direkten Schallanteils der Raumimpulsantwort und der Energie des Nachhalls einschließlich der frühen Reflexionen

$e_{\hat{T}_{60}}$ ..... Fehler in der Schätzung der Nachhallzeit

$\mathbf{e}_{m,k}^{(n)}$ ..... Fehler bei der Prädiktion eines Merkmalsvektors durch ein *SLDM*

$e_{\hat{\sigma}_h^2,\mathrm{REL}}$ ..... Relativer Schätzfehler in der Energie der Raumimpulsantwort

$\mathrm{EDC}_h(l)$ ..... Energieabfallkurve der Raumimpulsantwort

f_{A} ..... Abtastfrequenz

f_{O} ..... Nichtrekursive Beobachtungsfunktion

$\tilde{f}_{\mathrm{O}}$ ..... Vereinfachte nichtrekursive Beobachtungsfunktion (gültig bei Abwesenheit von Hintergrundstörungen)

$f_{\mathrm{O},L_{\mathrm{R}}}^{(\mathrm{R})}$ ..... Rekursive Beobachtungsfunktion mit der Rekursionslänge L_{R}

$\tilde{f}_{\mathrm{O},L_{\mathrm{R}}}^{(\mathrm{R})}$ ..... Vereinfachte rekursive Beobachtungsfunktion mit der Rekursionslänge L_{R} (gültig bei Abwesenheit von Hintergrundstörungen)

$h(l)$ ..... Zeitdiskrete Raumimpulsantwort

$h_{k,k'}(m'')$ ..... Kreuzbandfilter für $k \neq k'$ bzw. Band-zu-Band-Filter für $k = k'$ (definiert in (5.93))

$\tilde{h}_{k,k'}(l)$ ..... Hilfsfunktion zur anschaulichen Darstellung der Kreuzbandfilter (definiert in (5.97))

$h_{m,q}^{(\mathrm{s})}$ ..... Log-MEL-spektraler Koeffizient der Raumimpulsantwort

$\bar{h}_{m',q}$ ..... Koeffizient der Raumimpulsantwort im log-MEL-spektralen Bereich

$\bar{\mathbf{h}}_{m'}$ ..... Vektor der Koeffizient der Raumimpulsantwort im log-MEL-spektralen Bereich

$H(m,k)$ ..... Diskretes Kurzzeit-Spektrum der Raumimpulsantwort

$H\left(e^{j\theta}\right)$ ..... Zeitdiskrete FOURIER-Transformierte der Raumimpulsantwort

$H_{k,k'}\left(e^{j\theta}\right)$ ..... Zeitdiskrete FOURIER-Transformierte eines Kreuzbandfilters $h_{k,k'}(m'')$

$\tilde{H}_{k,k'}\left(e^{j\theta}\right)$ ..... Zeitdiskrete FOURIER-Transformierte von $\tilde{h}_{k,k'}(l)$

$\mathbf{H}_{f_{\mathrm{O}},\hat{\mathbf{n}}_{m|m,i}^{(\mathrm{s}),[r]}}$ ..... JACOBI-Matrix von f_{O} bezüglich $\mathbf{n}_m^{(\mathrm{s})}$ ausgewertet an der Stelle $\hat{\mathbf{n}}_{m|m,i}^{(\mathrm{s}),[r]}$ (genaue Definition in (5.218))

$\mathbf{H}_{f_{\mathrm{O}},\hat{\mathbf{x}}_m^{(\mathrm{s})}}$ ..... JACOBI-Matrix von f_{O} bezüglich $\mathbf{x}_m^{(\mathrm{s})}$ ausgewertet an der Stelle $\hat{\mathbf{x}}_m^{(\mathrm{s})}$ (genaue Definition in (5.219))

$\mathbf{H}_{f_{\mathrm{O}},\hat{\mathbf{z}}_{m|m,i}^{(\mathrm{s}),[r]}}$ ..... JACOBI-Matrix von f_{O} bezüglich $\mathbf{z}_m^{(\mathrm{s})}$ ausgewertet an der Stelle $\hat{\mathbf{z}}_{m|m,i}^{(\mathrm{s}),[r]}$ (genaue Definition in (5.216))

$\mathbf{H}_{f_{\mathrm{O}},\hat{\boldsymbol{\chi}}_{m|m,i}^{(\mathrm{s}),[r]}}$ ..... JACOBI-Matrix von f_{O} bezüglich $\boldsymbol{\chi}_m^{(\mathrm{s})}$ ausgewertet an der Stelle $\hat{\boldsymbol{\chi}}_{m|m,i}^{(\mathrm{s}),[r]}$ (genaue Definition in (5.217))

$\mathbf{H}_{f_{\mathrm{O},L_{\mathrm{C}}}^{(\mathrm{R})},\hat{\mathbf{n}}_m^{(\mathrm{s})}}$ ..... JACOBI-Matrix von $f_{\mathrm{O},L_{\mathrm{C}}-L_{\mathrm{C}}}^{(\mathrm{R})}$ bezüglich $\mathbf{n}_m^{(\mathrm{s})}$ ausgewertet an der Stelle $\hat{\mathbf{n}}_{m-L_{\mathrm{C}}}^{(\mathrm{s})}$ (genaue Definition in (5.225))

$\mathbf{H}_{f_{\mathrm{O},L_{\mathrm{C}}}^{(\mathrm{R})},\hat{\mathbf{n}}_{m|m,i}^{(\mathrm{s}),[r]}}$ .... JACOBI-Matrix von $f_{\mathrm{O},L_{\mathrm{C}}}^{(\mathrm{R})}$ bezüglich $\mathbf{n}_m^{(\mathrm{s})}$ ausgewertet an der Stelle $\hat{\mathbf{n}}_{m|m,i}^{(\mathrm{s}),[r]}$ (genaue Definition in (5.224))

$\mathbf{H}_{f_{\mathrm{O},L_\mathrm{C}}^{(\mathrm{R})},\hat{\mathbf{z}}_{m|m,i}^{(\mathrm{s}),[r]}}$ .... JACOBI-Matrix von $f_{\mathrm{O},L_\mathrm{C}}^{(\mathrm{R})}$ bezüglich $\mathbf{z}_m^{(\mathrm{s})}$ ausgewertet an der Stelle $\hat{\mathbf{z}}_{m|m,i}^{(\mathrm{s}),[r]}$ (genaue Definition in (5.222))

$\mathbf{H}_{f_{\mathrm{O},L_\mathrm{C}}^{(\mathrm{R})},\hat{\boldsymbol{\chi}}_{m|m,i}^{(\mathrm{s}),[r]}}$ ... JACOBI-Matrix von $f_{\mathrm{O},L_\mathrm{C}}^{(\mathrm{R})}$ bezüglich $\boldsymbol{\chi}_m^{(\mathrm{s})}$ ausgewertet an der Stelle $\hat{\boldsymbol{\chi}}_{m|m,i}^{(\mathrm{s}),[r]}$ (genaue Definition in (5.223))

$\mathbf{H}_{\hat{\mathbf{z}}_{m|m,i}^{(\mathrm{s}),[r]}}$ ..... JACOBI-Matrix der verwendeten Beobachtungsfunktion bezüglich $\mathbf{z}_m^{(\mathrm{s})}$ ausgewertet an der Stelle $\hat{\mathbf{z}}_{m|m,i}^{(\mathrm{s}),[r]}$

i ..... Index eines *SLDM*-Teilmodells

I ..... Anzahl der Teilmodelle eines *SLDM*

I_1 ..... Einseitige Länge des Fensters (in Anzahl von Segmenten) zur Berechnung der DELTA-Merkmale

I_2 ..... Einseitige Länge des Fensters (in Anzahl von Segmenten) zur Berechnung der DELTA-DELTA-Merkmale

$\mathscr{I}$ ..... Indexmenge aller wohl repräsentierten Teilmodelle eines *SLDM*

j ..... Imaginäre Einheit oder Index einzelner Experimente (aus dem Zusammenhang erkennbar)

J ..... Gesamtanzahl der Experimente

k ..... Frequenzindex

K ..... Anzahl der Frequenzbins (bei der *DFT* zur Merkmalsextraktion)

K' ..... Anzahl der cepstralen Koeffizienten (bei der Merkmalsextraktion)

K_q ..... Breite des q-ten MEL-Bandes (in Anzahl von Frequenzindizes)

$K_q^{(\mathrm{o})}$ ..... Obere Grenze des q-ten MEL-Bandes (in Form eines Frequenzindex)

$K_q^{(\mathrm{u})}$ ..... Untere Grenze des q-ten MEL-Bandes (in Form eines Frequenzindex)

$\mathbf{K}_{m,i}^{[r]}$ ..... KALMAN-Verstärkungsmatrix

l ..... Zeitindex (diskret) oder Index der *EM*-Iterationen (aus Zusammenhang erkennbar)

l_D ..... Zeitindex zur Bezeichnung des Zeitpunktes innerhalb der Raumimpulsantwort, an dem der Hauptimpuls auftritt

$\mathscr{L}(\boldsymbol{\theta})$ ..... Likelihoodfunktion

L_AR ..... Ordnung eines *SLDM*

L_C ..... Anzahl von aufeinanderfolgenden Merkmalsvektoren des sauberen Sprachsignals innerhalb des Zustandsvektors bei der KALMAN-Filterung

L_EM ..... Anzahl von *EM*-Iterationen

L_h ..... Länge der Raumimpulsantwort

L_H ..... Länge der Repräsentation der RIA im log-MEL-spektralen Bereich

L_R ..... Rekursionslänge für das rekursive Beobachtungsmodell

L_S ..... Länge der Merkmalsvektorsequenzen bei der *K-Means++*-artigen Initialisierung der *SLDM*-Parameter

L_w ..... Fensterlänge (bei der Merkmalsextraktion)

m ..... Segmentindex (diskret)

M ..... Anzahl von Merkmalsvektoren (bzw. Segmenten) innerhalb einer Sprachäußerung

M_n ..... Anzahl von Merkmalsvektoren innerhalb der n-ten Sprachäußerung

$\mathfrak{M}_{\mathrm{SEQ},k}(i)$ ..... Menge der zum k-ten Modell zugeordneten Merkmalsvektorsequenzen bei der Initialisierung der *SLDM*-Parameter (definiert in (5.70))

$\mathfrak{M}_{\text{SEQ},k,i}(I)$ ... Menge von Merkmalsvektorsequenztupeln (definiert in (5.75))
n ..... Index der Sprachäußerung innerhalb der Trainingsdaten
$n(l)$ ..... Zeitdiskretes Störsignal (nach der Versatzkompensation und der Höhenanhebung)
$\mathbf{n}_m^{(\text{s})}$ ..... Vektor der log-MEL-spektralen Koeffizienten des Störsignals
$\hat{\mathbf{n}}_m^{(\text{s})}$ ..... A-posteriori-Schätzwert des Vektors der log-MEL-spektralen Koeffizienten des Störsignals
$\hat{\mathbf{n}}_{m|m,i}^{(\text{s}),[r]}$ ..... Teilvektor von $\hat{\mathbf{z}}_{m|m,i}^{(\text{s}),[r]}$ bestehend aus der Schätzung des LMSK-Vektors des Störsignals
N ..... Anzahl der Sprachäußerungen innerhalb der Trainingsdaten
N_w ..... Anzahl von Wörtern innerhalb einer Sprachäußerung
$N^{(\text{SM})}$ ..... Anzahl (minus eins) vorhergehender Wörter, von denen ein Wort innerhalb eines Sprachmodells abhängig ist
$N(m,k)$ ..... Diskretes Kurzzeit-Spektrum des Störsignals
$\mathcal{N}_{m,q}$ ..... MEL-spektraler Koeffizient des Störsignals
N_{Einf} ..... Anzahl von Einfügefehlern bei der Spracherkennung
N_{Ausl} ..... Anzahl von Auslöschungsfehlern bei der Spracherkennung
N_{Subst} ..... Anzahl von Ersetzungsfehlern bei der Spracherkennung
N_{Ges} ..... Gesamtanzahl der Wörter innerhalb der Testdaten
P_k ..... Empirisch bestimmte Modellwahrscheinlickeiten
$P_{m|m-1,i}$ ..... A-priori-Modellwahrscheinlichkeiten (definiert in (5.230))
$P_{m|m,i}$ ..... A-posteriori-Modellwahrscheinlichkeiten (definiert in (5.232))
q ..... Index des MEL-Bandes
Q ..... Anzahl der MEL-Bänder (bei der Merkmalsextraktion)
$\mathcal{Q}_l(\theta)$ ..... Zu maximierende Hilfsfunktion beim *EM*-Algorithmus (definiert in (5.24))
r ..... Index der Iterationen beim *IEKF*
R ..... Anzahl der Iterationen beim *IEKF*
$s(l)$ ..... Zeitdiskretes verhalltes Sprachsignal (nach der Versatzkompensation und der Höhenanhebung)
$s_{m,q}^{(\text{s})}$ ..... Log-MEL-spektraler Koeffizient des verhallten Sprachsignals
$\hat{s}_{m,q}^{(\text{s})}$ ..... A-posteriori-Schätzwert des log-MEL-spektralen Koeffizienten des verhallten Sprachsignals
$\mathbf{s}_m^{(\text{s})}$ ..... Vektor der log-MEL-spektralen Koeffizienten des verhallten Sprachsignals
$\mathfrak{S}_i$ ..... Teilmenge der Parameter des i-ten Teilmodells eines *SLDM* (definiert in (5.58))
$S\left(e^{j\theta}\right)$ ..... Zeitdiskrete FOURIER-Transformierte des verhallten Sprachsignals
$S(m,k)$ ..... Diskretes Kurzzeit-Spektrum des verhallten Sprachsignals
t ..... Zeit (kontinuierlich)
T_{A} ..... Abtastdauer
T_{60} ..... Nachhallzeit
$\mathbf{U}_{\mathbf{V}_i}$ ..... Eigenvektormatrix von $\mathbf{V}_i$
$\mathbf{U}_{\boldsymbol{\Sigma}_{\mathbf{x},i}}$ ..... Eigenvektormatrix von $\boldsymbol{\Sigma}_{\mathbf{x},i}$

$v_h(l)$ ..... Zeitdiskreter weißer GAUSS'scher Zufallsprozess zur Erzeugung der Raumimpulsantwort gemäß einem vereinfachten Modell

$v_{m,q}^{(\mathrm{s})}$ ..... Beobachtungsfehler beim nichtrekursiven Beobachtungsmodell

$v_{m,q,L_\mathrm{R}}^{(\mathrm{s,R})}$ ..... Beobachtungsfehler beim rekursiven Beobachtungsmodell mit der Rekursionslänge L_R

$\mathbf{v}_m^{(\mathrm{s})}$ ..... Vektor der Beobachtungsfehler beim nichtrekursiven Beobachtungsmodell

$\hat{\mathbf{v}}_m^{(\mathrm{s})}$ ..... Vektor der approximativen Beobachtungsfehler beim nichtrekursiven Beobachtungsmodell unter Berücksichtigung von Modellunzulänglichkeiten und Schätzfehler in den Modellparametern

$\mathbf{v}_{m,L_\mathrm{R}}^{(\mathrm{s,R})}$ ..... Vektor der Beobachtungsfehler beim rekursiven Beobachtungsmodell mit der Rekursionslänge L_R

$\mathbf{V}_i$ ..... Kovarianzmatrix des Prädiktionsfehlers durch das i-te Teilmodell eines *SLDM*

$\mathbf{V}_i^{\{l\}}$ ..... Kovarianzmatrix des Prädiktionsfehlers durch das i-te Teilmodell eines *SLDM* berechnet nach der l-ten *EM*-Iteration

$w(l)$ ..... Zeitdiskretes Fenster entstehend aus der Faltung des Analysefensters $w_\mathrm{A}(l)$ mit dem zeitumgekehrten Synthesefenster $w_\mathrm{S}(-l)$

$w_\mathrm{A}(l)$ ..... Zeitdiskretes Analysefenster

$w_{\mathrm{MA},k}(l)$ ..... Zeitdiskretes und zeitumgekehrtes, moduliertes Analysefenster

$w_\mathrm{S}(l)$ ..... Zeitdiskretes Synthesefenster

$w_{\mathrm{MS},k}(l)$ ..... Zeitdiskretes moduliertes Synthesefenster

w_ν ..... ν-tes Wort innerhalb einer Sprachäußerung

$W_\mathrm{A}\left(e^{j\theta}\right)$ ..... Zeitdiskrete FOURIER-Transformierte des Analysefensters

$W_\mathrm{S}\left(e^{j\theta}\right)$ ..... Zeitdiskrete FOURIER-Transformierte des Synthesefensters

$x(l)$ ..... Zeitdiskretes sauberes Sprachsignal (nach der Versatzkompensation und der Höhenanhebung)

$\mathbf{x}_m^{(\mathrm{s})}$ ..... Vektor der log-MEL-spektralen Koeffizienten des sauberen Sprachsignals

$\mathbf{x}_m^{(n)}$ ..... Zum Training eines *SLDM* verwendeter Merkmalsvektor zugehörig zum Segment m der n-ten Sprachäußerung

$\mathbf{x}_m$ ..... Merkmalsvektor zusammengesetzt aus den cepstralen Koeffizienten und den DELTA- und DELTA-DELTA-Merkmalen des sauberen Sprachsignals

$\hat{\mathbf{x}}_m^{(\mathrm{s})}$ ..... A-posteriori-Schätzwert des Vektors der log-MEL-spektralen Koeffizienten des sauberen Sprachsignals

$\mathfrak{X}$ ..... Menge der Merkmalsvektorsequenzen aller Sprachäußerungen innerhalb der Trainingsdaten

$\mathfrak{X}_{1:L_\mathrm{AR}}$ ..... Menge der L_AR ersten Merkmalsvektoren aller Sprachäußerungen innerhalb der Trainingsdaten

$\mathfrak{X}_{\mathrm{SEQ},L_\mathrm{S}}$ ..... Menge aller möglichen Merkmalsvektorsequenzen innerhalb der Trainingsdaten

$X\left(e^{j\theta}\right)$ ..... Zeitdiskrete FOURIER-Transformierte des sauberen Sprachsignals

$X(m,k)$ ..... Diskretes Kurzzeit-Spektrum des sauberen Sprachsignals

$\mathscr{X}_{m,q}$ ..... MEL-spektraler Koeffizient des sauberen Sprachsignals

$y(l)$ ..... Zeitdiskretes verhalltes und gestörtes Sprachsignal (nach der Versatzkompensation und der Höhenanhebung)

$y_{\mathrm{MIC}}(l)$ ..... Zeitdiskretes (verhalltes und gestörtes) Mikrofonsignal (nach der Versatzkompensation und der Höhenanhebung)

$y_{w_{\mathrm{A}}}(m,l')$ ..... Gefenstertes zeitdiskretes verhalltes und gestörtes Sprachsignal

$y^{(\mathrm{c})}_{m,\kappa'}$ ..... Cepstraler Koeffizient des verhallten und gestörten Sprachsignals

$y^{(\mathrm{s})}_{m,q}$ ..... Log-MEL-spektraler Koeffizient des verhallten und gestörten Sprachsignals

$\hat{y}^{(\mathrm{s})}_{m,q}$ ..... A-posteriori-Schätzwert des log-MEL-spektralen Koeffizienten des verhallten und gestörten Sprachsignals

$\mathbf{y}_m$ ..... Merkmalsvektor zusammengesetzt aus den cepstralen Koeffizienten und ihren DELTA- und DELTA-DELTA-Merkmalen des verhallten und gestörten Sprachsignals

$\mathbf{y}^{(\mathrm{s})}_m$ ..... Vektor der log-MEL-spektralen Koeffizienten des verhallten und gestörten Sprachsignals

$\hat{\mathbf{y}}^{(\mathrm{s}),[r]}_{m,i}$ ..... Prädiktion für den beobachteten LMSK-Vektor des verhallten und gestörten Sprachsignals beruhend auf der Linearierungsstelle $\hat{\mathbf{z}}^{(\mathrm{s}),[r]}_{m|m,i}$

$\hat{\mathbf{y}}^{(\mathrm{s}),[r]}_{m,i,k}$ ..... Prädiktion für den beobachteten LMSK-Vektor des verhallten und gestörten Sprachsignals beruhend auf der Linearierungsstelle $\hat{\mathbf{z}}^{(\mathrm{s}),[r]}_{m|m,i,k}$

$Y\left(e^{j\theta}\right)$ ..... Zeitdiskrete FOURIER-Transformierte des verhallten und gestörten Sprachsignals

$Y(m,k)$ ..... Diskretes Kurzzeit-Spektrum des verhallten und gestörten Sprachsignals

$\mathscr{Y}_{m,q}$ ..... MEL-spektraler Koeffizient des verhallten und gestörten Sprachsignals

$\mathfrak{Z}$ ..... Menge der Zustandssequenzen aller Sprachäußerungen innerhalb der Trainingsdaten

$\mathbf{z}^{(\mathrm{s})}_m$ ..... Zusammengesetzter Vektor bestehend aus $\boldsymbol{\chi}^{(\mathrm{s})}_m$ und $\mathbf{n}^{(\mathrm{s})}_m$

$\hat{\mathbf{z}}^{(\mathrm{s})}_{m|m-1}$ ..... Geschätzter Mittelwertvektor beruhend auf der prädiktiven Verteilungsdichtefunktion $p\left(\mathbf{z}^{(\mathrm{s})}_m \middle| \mathbf{y}^{(\mathrm{s})}_{1:m-1}\right)$

$\hat{\mathbf{z}}^{(\mathrm{s})}_{m|m}$ ..... Geschätzter Mittelwertvektor beruhend auf der A-posteriori-Verteilungsdichtefunktion $p\left(\mathbf{z}^{(\mathrm{s})}_m \middle| \mathbf{y}^{(\mathrm{s})}_{1:m}\right)$

$\hat{\mathbf{z}}^{(\mathrm{s})}_{m|m-1,i}$ ..... Geschätzter Mittelwertvektor beruhend auf der prädiktiven Verteilungsdichtefunktion $p\left(\mathbf{z}^{(\mathrm{s})}_m \middle| \mathbf{y}^{(\mathrm{s})}_{1:m-1}, \zeta_m = i\right)$

$\hat{\mathbf{z}}^{(\mathrm{s})}_{m|m,i}$ ..... Geschätzter Mittelwertvektor beruhend auf der A-posteriori-Verteilungsdichtefunktion $p\left(\mathbf{z}^{(\mathrm{s})}_m \middle| \mathbf{y}^{(\mathrm{s})}_{1:m}, \zeta_m = i\right)$

$\hat{\mathbf{z}}^{(\mathrm{s})}_{m|m,i,k}$ ..... Geschätzter Mittelwertvektor beruhend auf der A-posteriori-Verteilungsdichtefunktion $p\left(\mathbf{z}^{(\mathrm{s})}_m \middle| \mathbf{y}^{(\mathrm{s})}_{1:m}, \zeta_{m-1} = i, \zeta_m = k\right)$

$\hat{\mathbf{z}}^{(\mathrm{s}),[r]}_{m|m,i}$ ..... Linearisierungsstelle der Beobachtungsfunktion bei der r-ten Iteration des *IEKF* zur Berechnung von $\hat{\mathbf{z}}^{(\mathrm{s})}_{m|m,i}$

$\hat{\mathbf{z}}^{(\mathrm{s}),[r]}_{m|m,i,k}$ ..... Linearisierungsstelle der Beobachtungsfunktion bei der r-ten Iteration des *IEKF* zur Berechnung von $\hat{\mathbf{z}}^{(\mathrm{s})}_{m|m,i,k}$

Griechische Formelzeichen

$\alpha^{(\mathrm{SM})}$ ..... Konstante zur Skalierung des Gewichtes des Sprachmodells gegenüber dem des akustischen Modells

$\alpha^{(n,l)}_m(i)$ ..... Vorwärtswahrscheinlichkeit (definiert in (A.13))

α_h ..... Negativer Exponent von ε_h zur Basis 10

β ..... Skalierungsfaktor zur Festlegung des Ausmaßes der Pertubation bei der Modellspaltung

$\beta^{(n,l)}_m(i)$ ..... Rückwärtswahrscheinlichkeit (definiert int (A.14))

$\gamma_{1:M}$ ..... Sequenz der Zustände innerhalb eines HMM

$\delta^{(l)}_{\mathscr{L}}$ ..... Mittlere relative Verbesserung der Likelihoodfunktion pro einzelne Äußerung (definiert in (5.38))

$\Delta_{\mathfrak{X}}$ ..... Menge bestehend aus den Differenzen aufeinanderfolgender Merkmalsvektoren aller Sprachäußerungen innerhalb der Trainingsdaten

$\Delta y^{(\mathrm{c})}_{m,\kappa'}$ ..... DELTA-Merkmal des verhallten und gestörten Sprachsignals

$\Delta\Delta y^{(\mathrm{c})}_{m,\kappa'}$ ..... DELTA-DELTA-Merkmal des verhallten und gestörten Sprachsignals

ε_h ..... Konstante zur Festlegung des maximalen relativen Fehlers in der Energie der Raumimpulsantwort, der durch zeitliches Abschneiden eingeführt wird

$\varepsilon_{\mathscr{L}}$ ..... Untere Schranke für die mittlere relative Verbesserung der Likelihoodfunktion pro einzelne Äußerung

$\varepsilon_{P,\mathrm{REL}}$ ..... Konstante, die angibt, wie zahlreich ein Teilmodell bei der *SLDM*-Initialisierung mindestens im Verhältnis zum bestrepräsentierten Teilmodell vertreten sein sollte

ζ_m ..... Aktives Teilmodell innerhalb eines *SLDM* zum Segmentindex m

$\zeta^{(n)}_m$ ..... Aktives Teilmodell innerhalb eines *SLDM* zum Segmentindex m der n-ten Sprachäußerung

$\eta^{(n,l)}_m(i)$ ..... Bedingte Wahrscheinlichkeit für die Aktivität des i-ten Teilmodells eines *SLDM* (definiert in (5.26))

$\Omega^{(n)}_{\mathrm{SEQ},m}(i)$ ..... Zugehörigkeit der Merkmalsvektorsequenz einer Sprachäußerung zu einem Teilmodell eines *SLDM* (definiert in (5.71))

θ ..... Normierte Kreisfrequenz

θ ..... Menge aller Parameter eines *SLDM*

κ' ..... Index der cepstralen Koeffizienten

λ_{Ausl} ..... Rate der Auslöschungsfehler bei der Spracherkennung

λ_{Einf} ..... Rate der Einfügefehler bei der Spracherkennung

λ_{Subst} ..... Rate der Ersetzungsfehler bei der Spracherkennung

λ_w ..... Wortfehlerrate bei der Spracherkennung

$\boldsymbol{\Lambda}_{\mathbf{V}_i}$ ..... Eigenwertmatrix von $\mathbf{V}_i$

$\boldsymbol{\Lambda}_{\boldsymbol{\Sigma}_{\mathbf{x},i}}$ ..... Eigenwertmatrix von $\boldsymbol{\Sigma}_{\mathbf{x},i}$

$\mu_{\breve{h}_{m',q}}$ ..... Mittelwert des Koeffizienten der Raumimpulsantwort im log-MEL-spektralen Bereich beruhend auf dem Modell der Raumimpulsantwort

$\boldsymbol{\mu}_{\breve{\mathbf{h}}_{m'}}$ ..... Vektor der Mittelwerte der Koeffizienten der Raumimpulsantwort im log-MEL-spektralen Bereich beruhend auf dem Modell der Raumimpulsantwort

$\boldsymbol{\mu}_{\mathbf{n}}$ ..... Mittelwertvektor für das A-priori-Modell für die LMSK-Vektoren des Störsignals

$\boldsymbol{\mu}_{\mathbf{x},i}$ ..... Mittelwertvektor der i-ten Mischungskomponente des *GMM* zur Modellierung der ersten L_{AR} LMSK-Vektoren des sauberen Sprachsignals innerhalb einer Sprachäußerung

$\boldsymbol{\mu}_{\mathbf{x},i}^{\{l\}}$ ..... Mittelwertvektor der i-ten Mischungskomponente des *GMM* zur Modellierung der ersten L_{AR} LMSK-Vektoren des sauberen Sprachsignals innerhalb einer Sprachäußerung berechnet nach der l-ten *EM*-Iteration

$\boldsymbol{\mu}_{\breve{\mathbf{x}}_m^{(\mathrm{s})}\left|\mathbf{y}_{1:m}^{(\mathrm{s})}\right.}$ ..... Mittelwert von $\breve{\mathbf{x}}_m^{(\mathrm{s})}$ bedingt auf die Beobachtung von $\breve{\mathbf{y}}_{1:m}^{(\mathrm{s})}$

$\boldsymbol{\mu}_{\hat{\mathbf{v}}^{(\mathrm{s})}}$ ..... Mittelwertvektor des approximativen Beobachtungsfehlervektors beim nichtrekursiven Beobachtungsmodell

$\boldsymbol{\mu}_{\hat{\mathbf{v}}_{L_{\mathrm{R}}}^{(\mathrm{s,R})}}$ ..... Mittelwertvektor des approximativen Beobachtungsfehlervektors beim rekursiven Beobachtungsmodell mit der Rekursionslänge L_{R}

$\xi_m^{(n,l)}(k,i)$ ..... Bedingte Wahrscheinlichkeit für die aufeinanderfolgende Aktivität zweier Teilmodelle eines *SLDM* (definiert in (5.27))

$\rho_{\hat{\boldsymbol{\Sigma}}_{\hat{\mathbf{v}}^{(\mathrm{s})}}}$ ..... Spektralradius von $\boldsymbol{\Sigma}_{\hat{\mathbf{v}}^{(\mathrm{s})}}$

$\sigma_{m,k}$ ..... GABOR-Koeffizient zur Darstellung des verhallten Sprachsignals $s(l)$

$\sigma_{\breve{h}_{m',q}}^2$ ..... Varianz des Koeffizienten der Raumimpulsantwort im log-MEL-spektralen Bereich beruhend auf dem Modell der Raumimpulsantwort

σ_h^2 ..... Energie der Raumimpulsantwort

σ_n^2 ..... Leistung des Störsignals

σ_s^2 ..... Leistung des verhallten Sprachsignals

σ_x^2 ..... Leistung des sauberen Sprachsignals

σ_y^2 ..... Leistung des verhallten und gestörten Sprachsignals

$\boldsymbol{\Sigma}_{\mathbf{n}}$ ..... Kovarianzmatrix für das A-priori-Modell für die LMSK-Vektoren des Störsignals

$\hat{\boldsymbol{\Sigma}}_{\breve{\mathbf{n}}_m^{(\mathrm{s})}}$ ..... Approximative Schätzfehlerkovarianzmatrix für die Schätzung des Vektors der log-MEL-spektralen Koeffizienten des Störsignals

$\boldsymbol{\Sigma}_{\mathbf{x},i}$ ..... Kovarianzmatrix der i-ten Mischungskomponente des *GMM* zur Modellierung der ersten L_{AR} LMSK-Vektoren des sauberen Sprachsignals innerhalb einer Sprachäußerung

$\boldsymbol{\Sigma}_{\mathbf{x},i}^{\{l\}}$ ..... Kovarianzmatrix der i-ten Mischungskomponente des *GMM* zur Modellierung der ersten L_{AR} LMSK-Vektoren des sauberen Sprachsignals innerhalb einer Sprachäußerung berechnet nach der l-ten *EM*-Iteration

$\boldsymbol{\Sigma}_{\breve{\mathbf{x}}_m^{(\mathrm{s})}\left|\mathbf{y}_{1:m}^{(\mathrm{s})}\right.}$ ..... Kovarianzmatrix von $\breve{\mathbf{x}}_m^{(\mathrm{s})}$ bedingt auf die Beobachtung von $\breve{\mathbf{y}}_{1:m}^{(\mathrm{s})}$

$\hat{\boldsymbol{\Sigma}}_{\hat{\mathbf{x}}_m^{(\mathrm{s})}}$ ..... Approximative Schätzfehlerkovarianzmatrix für die Schätzung des Vektors der log-MEL-spektralen Koeffizienten des sauberen Sprachsignals

$\boldsymbol{\Sigma}_{\hat{\mathbf{v}}^{(\mathrm{s})}}$ ..... Kovarianzmatrix des approximativen Beobachtungsfehlervektors beim nichtrekursiven Beobachtungsmodell

$\boldsymbol{\Sigma}_{\hat{\mathbf{v}}_{L_\mathrm{R}}^{(\mathrm{s,R})}}$ ..... Kovarianzmatrix des approximativen Beobachtungsfehlervektors beim rekursiven Beobachtungsmodell mit der Rekursionslänge L_R

$\hat{\boldsymbol{\Sigma}}_{\mathbf{y}_{m,i}^{(\mathrm{s}),[r]}}$ ..... Kovarianzmatrix der Prädiktion $\hat{\mathbf{y}}_{m,i}^{(\mathrm{s}),[r]}$ für den beobachteten LMSK-Vektor des verhallten und gestörten Sprachsignals

$\hat{\boldsymbol{\Sigma}}_{\mathbf{y}_{m,i,k}^{(\mathrm{s}),[r]}}$ ..... Kovarianzmatrix der Prädiktion $\hat{\mathbf{y}}_{m,i,k}^{(\mathrm{s}),[r]}$ für den beobachteten LMSK-Vektor des verhallten und gestörten Sprachsignals

$\hat{\boldsymbol{\Sigma}}_{\mathbf{z}_{m|m-1}^{(\mathrm{s})}}$ ..... Geschätzte Kovarianzmatrix beruhend auf der prädiktiven Verteilungsdichtefunktion $p\left(\mathbf{z}_m^{(\mathrm{s})} \middle| \mathbf{y}_{1:m-1}^{(\mathrm{s})}\right)$

$\hat{\boldsymbol{\Sigma}}_{\mathbf{z}_{m|m}^{(\mathrm{s})}}$ ..... Geschätzte Kovarianzmatrix beruhend auf der A-posteriori-Verteilungsdichtefunktion $p\left(\mathbf{z}_m^{(\mathrm{s})} \middle| \mathbf{y}_{1:m}^{(\mathrm{s})}\right)$

$\hat{\boldsymbol{\Sigma}}_{\mathbf{z}_{m|m-1,i}^{(\mathrm{s})}}$ ..... Geschätzte Kovarianzmatrix beruhend auf der prädiktiven Verteilungsdichtefunktion $p\left(\mathbf{z}_m^{(\mathrm{s})} \middle| \mathbf{y}_{1:m-1}^{(\mathrm{s})}, \zeta_m = i\right)$

$\hat{\boldsymbol{\Sigma}}_{\mathbf{z}_{m|m,i}^{(\mathrm{s})}}$ ..... Geschätzte Kovarianzmatrix beruhend auf der A-posteriori-Verteilungsdichtefunktion $p\left(\mathbf{z}_m^{(\mathrm{s})} \middle| \mathbf{y}_{1:m}^{(\mathrm{s})}, \zeta_m = i\right)$

$\hat{\boldsymbol{\Sigma}}_{\mathbf{z}_{m|m,i,k}^{(\mathrm{s})}}$ ..... Geschätzte Kovarianzmatrix beruhend auf der A-posteriori-Verteilungsdichtefunktion $p\left(\mathbf{z}_m^{(\mathrm{s})} \middle| \mathbf{y}_{1:m}^{(\mathrm{s})}, \zeta_{m-1} = i, \zeta_m = k\right)$

τ_h ..... Abklingkonstante der Raumimpulsantwort

$\phi_{k,k'}(l)$ ..... Zeitdiskrete Hilfsfunktion zur vereinfachten Darstellung der Kreuzbandfilter (definiert in (5.96))

$\Phi_{k,k'}\left(e^{j\theta}\right)$ ..... Zeitdiskrete FOURIER-Transformierte von $\phi_{k,k'}(l)$

$\chi_h(l)$ ..... Zeitdiskrete Indikatorfunktion der Raumimpulsantwort

$\boldsymbol{\chi}_m^{(\mathrm{s})}$ ..... Zusammengesetzter Vektor bestehend aus L_C zeitlich aufeinanderfolgenden Vektoren der log-MEL-spektralen Koeffizienten des sauberen Sprachsignals, d.h. $\mathbf{x}_m^{(\mathrm{s})}, \ldots, \mathbf{x}_{m-L_\mathrm{C}+1}^{(\mathrm{s})}$

$\hat{\boldsymbol{\chi}}_{m|m,i}^{(\mathrm{s}),[r]}$ ..... Teilvektor von $\hat{\mathbf{z}}_{m|m,i}^{(\mathrm{s}),[r]}$ bestehend aus Schätzungen von L_C aufeinanderfolgenden LMSK-Vektoren des sauberen Sprachsignals

ψ_i ..... Wahrscheinlichkeit für die Aktivität des i-ten Teilmodells des *SLDM* innerhalb der ersten L_AR Merkmalsvektoren einer Sprachäußerung

$\psi_i^{\{l\}}$ ..... Wahrscheinlichkeit für die Aktivität des i-ten Teilmodells des *SLDM* innerhalb der ersten L_AR Merkmalsvektoren einer Sprachäußerung berechnet nach der l-ten *EM*-Iteration

Abbildungsverzeichnis

Tabellenverzeichnis

Literaturverzeichnis

[AB57] J. Aitchison und J. A. C. Brown: *The Lognormal Distribution: with Special Reference to its Uses in Economics*, Cambridge University Press, Cambridge, 1957.

[AC07a] Y. Avargel und I. Cohen: „On Multiplicative Transfer Function Approximation in the Short-Time Fourier Transform Domain“, *IEEE Signal Processing Letters*, Band 14(5), S. 337–340, Mai 2007.

[AC07b] Y. Avargel und I. Cohen: „System Identification in the Short-Time Fourier Transform Domain With Crossband Filtering“, *IEEE Transactions on Audio, Speech, and Language Processing*, Band 15(4), S. 1305–1319, Mai 2007.

[All79] J. B. Allen: „Image Method for Efficiently Simulating Small-Room Acoustics“, *The Journal of the Acoustical Society of America*, Band 65(4), S. 943–950, Apr. 1979.

[AMGC02] M. Arulampalam, S. Maskell, N. Gordon und T. Clapp: „A Tutorial on Particle Filters for Online Nonlinear/Non-Gaussian Bayesian Tracking“, *IEEE Transactions on Signal Processing*, Band 50(2), S. 174–188, Febr. 2002.

[Ata95] B. S. Atal: „Speech Technology in 2001: New Research Directions“, *Proceedings of the National Academy of Sciences of the United States of America*, Band 92(22), S. 10046–10051, Okt. 1995.

[ATH97] C. Avendano, S. Tibrewala und H. Hermansky: „Multiresolution Channel Normalization for ASR in Reverberant Environments“, *Proc. of European Conference on Speech Communication and Technology (EUROSPEECH)*, S. 1107–1110, Rhodes, Greece, Sept. 1997.

[AV07] D. Arthur und S. Vassilvitskii: „K-Means++: the Advantages of Careful Seeding“, *Proc. of Symposium on Discrete Algorithms (SODA)*, S. 1027–1035, 2007.

[Ave97] C. Avendano: *Temporal Processing of Speech in a Time-Feature Space*, Dissertation, Oregon Graduate Institute of Science & Technology, 1997.

[BD86] G. E. P. Box und N. R. Draper: *Empirical Model-Building and Response Surface*, John Wiley & Sons, Inc., New York, NY, USA, 1986.

[BSLK01] Y. Bar-Shalom, X. R. Li und T. Kirubarajan: *Estimation with Applications to Tracking and Navigation: Theory, Algorithms, and Software*, Wiley, New York, 2001.

[CB07] C.-P. Chen und J. A. Bilmes: „MVA Processing of Speech Features“, *IEEE Transactions on Audio, Speech, and Language Processing*, Band 15(1), S. 257–270, 2007.

[CC04] L. Couvreur und C. Couvreur: „Blind Model Selection for Automatic Speech Recognition in Reverberant Environments“, Band 36(2/3), S. 189–203, 2004.

[CGJV01] M. Cooke, P. Green, L. Josifovski und A. Vizinho: „Robust Automatic Speech Recognition with Missing and Unreliable Acoustic Data“, *Speech Communication*, Band 34(3), S. 267–285, 2001.

[CR83] R. E. Crochiere und L. R. Rabiner: *Multirate Digital Signal Processing*, Prentice Hall, 1983.

[CZ98] R. A. Cole und V. Zue: „Spoken Language Input“, R. A. Cole, J. Mariani, H. Uszkoreit, A. Zaenen, V. Zue, G. Varile und A. Zampolli, Hrsg., *Survey of the State of the Art in Human Language Technology (Studies in Natural Language Processing)*, S. 1–62, Cambridge University Press, 1998.

[DBY07] J. Deng, M. Bouchard und T. H. Yeap: „Noisy Speech Feature Estimation on the Aurora2 Database Using a Switching Linear Dynamic Model“, *Journal of Multimedia*, Band 2(2), S. 47–52, 2007.

[DHM07] M. Delcroix, T. Hikichi und M. Miyoshi: „Precise Dereverberation Using Multichannel Linear Prediction“, *IEEE Transactions on Audio, Speech, and Language Processing*, Band 15(2), S. 430–440, Febr. 2007.

[DHS01] R. O. Duda, P. E. Hart und D. G. Stork: *Pattern Classification*, Wiley-Interscience, 2. Aufl., Nov. 2001.

[DLR77] A. P. Dempster, N. M. Laird und D. B. Rubin: „Maximum Likelihood from Incomplete Data via the EM Algorithm“, *Journal of the Royal Statistical Society. Series B (Methodological)*, Band 39(1), S. 1–38, 1977.

[dlTPS+05] A. de la Torre, A. Peinado, J. Segura, J. Perez-Cordoba, M. Benitez und A. Rubio: „Histogram Equalization of Speech Representation for Robust Speech Recognition“, *IEEE Transactions on Speech and Audio Processing*, Band 13(3), S. 355–366, Mai 2005.

[DM80] S. Davis und P. Mermelstein: „Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences“, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Band 28(4), S. 357–366, Aug. 1980.

[DNW09] M. Delcroix, T. Nakatani und S. Watanabe: „Static and Dynamic Variance Compensation for Recognition of Reverberant Speech With Dereverberation Preprocessing“, *IEEE Transactions on Audio, Speech, and Language Processing*, Band 17(2), S. 324–334, Febr. 2009.

[ETSa] ETSI: *ETSI standard document, Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms, ETSI ES 202 050 V1.1.5 (2007-01).*

[ETSb] ETSI: *ETSI standard document, Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithms, ETSI ES 201 108 V1.1.3 (2003-09).*

[FJZE85] J. L. Flanagan, J. D. Johnston, R. Zahn und G. W. Elko: „Computer-Steered Microphone Arrays for Sound Transduction in Large Rooms“, *The Journal of the Acoustical Society of America*, Band 78(5), S. 1508–1518, 1985.

[FR94] S. Farkash und S. Raz: „Linear Systems in Gabor Time-Frequency Space“, *IEEE Transactions on Signal Processing*, Band 42(3), S. 611–617, März 1994.

[Fur81] S. Furui: „Cepstral Analysis Technique for Automatic Speaker Verification“, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Band 29(2), S. 254–272, 1981.

[Gal98] M. J. F. Gales: „Maximum Likelihood Linear Transformations for HMM-Based Speech Recognition“, Band 12(2), S. 75–98, 1998.

[Gan08] S. Gannot: „Multi-Microphone Speech Dereverberation Based on Eigen-Decomposition: A Study“, *Proc. of Asilomar Conference on Signals, Systems and Computers (ACSSC)*, S. 801–805, Pacific Grove, CA, USA, Okt. 2008.

[Gan10] S. Gannot: „Multi-Microphone Speech Dereverberation Using Eigen-Decomposition“, P. A. N. und Nikolay D. Gaubitch, Hrsg., *Speech Dereverberation*, Kap. 5, Springer, 2010.

[GC89] L. Gillick und S. Cox: „Some Statistical Issues in the Comparison of Speech Recognition Algorithms“, *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Band 1, S. 532–535, Mai 1989.

[GD97] D. Gesbert und P. Duhamel: „Robust Blind Channel Identification and Equalization Based on Multi-Step Predictors“, *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, S. 3621–3624, Munich, Germany, Apr. 1997.

[GK97] S. Greenberg und B. Kingsbury: „The Modulation Spectrogram: In Pursuit of an Invariant Representation of Speech“, *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, S. 1647–1650, Munich, Germany, Apr. 1997.

[GM01] D. Gelbart und N. Morgan: „Evaluating Long-Term Spectral Subtraction for Reverberant ASR“, *Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, S. 103–106, Madonna Di Campiglio, Italy, Dez. 2001.

[GM02] D. Gelbart und N. Morgan: „Double the Trouble: Handling Noise and Reverberation in Far-Field Automatic Speech Recognition“, *Proc. of International Conference on Spoken Language Processing (ICSLP)*, S. 2185–2188, Denver, CO, USA, Sept. 2002.

[GM03] S. Gannot und M. Moonen: „Subspace Methods for Multi-Microphone Speech Dereverberation“, Band 11, S. 1074–1090, 2003.

[GMF01] B. Gillespie, H. Malvar und D. Florencio: „Speech Dereverberation via Maximum-Kurtosis Subband Adaptive Filtering“, *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, S. 3701–3704, Salt Lake City, UT, USA, Mai 2001.

[GMOS99] D. Giuliani, M. Matassoni, M. Omologo und P. Svaizer: „Training of HMM with Filtered Speech Material for Hands-Free Recognition“, *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, S. 449–452, Phoenix, AZ, USA, März 1999.

[GN95] M. Gürelli und C. Nikias: „EVAM: An Eigenvector-Based Algorithm for Multichannel Blind Deconvolution of Input Colored Signals“, *IEEE Transactions on Signal Processing*, Band 43(1), S. 134–149, Jan. 1995.

[GNW03] N. Gaubitch, P. Naylor und D. Ward: „On the Use of Linear Prediction for Dereverberation of Speech“, *Proc. of International Workshop on Acoustic Echo and Noise Control (IWAENC)*, S. 99–102, Kyoto, Japan, Sept. 2003.

[Gol67] J. L. Goldstein: „Auditory Spectral Filtering and Monaural Phase Perception“, *The Journal of the Acoustical Society of America*, Band 41(2), S. 458–479, 1967.

[GPAF04] S. Greenberg, A. N. Popper, W. A. Ainsworth und R. R. Fay: *Speech Processing in the Auditory System*, Springer Verlag, 2004.

[Gre61] D. D. Greenwood: „Critical Bandwidth and the Frequency Coordinates of the Basilar Membrane“, *The Journal of the Acoustical Society of America*, Band 33(10), S. 1344–1356, 1961.

[GRTN10] N. D. Gaubitch, M. R, P. Thomas und P. A. Naylor: „Dereverberation Using LPC-based Approaches“, *Speech Dereverberation*, Kap. 4, S. 95–128, Springer, 2010.

[GW96] M. J. F. Gales und P. C. Woodland: „Mean and Variance Adaptation within the MLLR Framework“, Band 10(4), S. 249–264, 1996.

[GY95] M. F. J. Gales und S. J. Young: „Robust Speech Recognition in Additive and Convolutional Noise Using Parallel Model Combination“, *Computer Speech & Language*, Band 9(4), S. 289–307, 1995.

[Hab04] E. A. P. Habets: „Single-Channel Speech Dereverberation Based on Spectral Subtraction“, *Proc. of Annual Workshop on Circuits, Systems and Signal Processing (ProRISC)*, S. 250–254, Veldhoven, The Netherlands, Nov. 2004.

[Hab07] E. Habets: *Single- and Multi-Microphone Speech Dereverberation Using Spectral Enhancement*, Dissertation, Technische Universiteit Eindhoven, Juni 2007.

[HBC08] Y. A. Huang, J. Benesty und J. Chen: „Dereverberation“, J. Benesty, M. M. Sondhi und Y. A. Huang, Hrsg., *Springer Handbook of Speech Processing*, S. 929–944, Springer Berlin Heidelberg, 2008.

[HDM06] T. Hikichi, M. Delcroix und M. Miyoshi: „On Robust Inverse Filtering Design For Room Transfer Function Fluctuations“, *Proc. of European Signal Processing Conference (EUSIPCO)*, Florence, Italy, Sept. 2006.

[HE95] H. G. Hirsch und C. Ehrlicher: „Noise Estimation Techniques for Robust Speech Recognition“, *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, S. 153–156, Detroit, MI, USA, 1995.

[Her90] H. Hermansky: „Perceptual Linear Predictive (PLP) Analysis of Speech“, *The Journal of the Acoustical Society of America*, Band 87(4), S. 1738–1752, 1990.

[Her96] J. D. Herre, Jürgen; Johnston: „Enhancing the Performance of Perceptual Audio Coders by Using Temporal Noise Shaping (TNS)“, *Proc. of Audio Engineering Society (AES) Convention*, Los Angeles, CA, USA, Nov. 1996.

[HF05] H.-G. Hirsch und H. Finster: „The Simulation of Realistic Acoustic Input Scenarios for Speech Recognition Systems“, *Proc. of Annual Conference of the International Speech Communication Association (Interspeech)*, S. 2697–2700, Lisbon, Portugal, Sept. 2005.

[HF08] H.-G. Hirsch und H. Finster: „A New Approach for the Adaptation of HMMs to Reverberation and Background Noise“, *Speech Communication*, Band 50(3), S. 244–263, 2008.

[HGH06] H. F. Hans-Günter Hirsch: „A New HMM Adaptation Approach for the Case of a Hands-Free Speech Input in Reverberant Rooms“, *Proc. of Annual Conference of the International Speech Communication Association (Interspeech)*, Pittsburgh, PA, USA, Sept. 2006.

[HHW85] H. Hermansky, B. Hanson und H. Wakita: „Perceptually Based Linear Predictive Analysis of Speech“, *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, S. 509–512, Tampa, FL, USA, Apr. 1985.

[Hir07] H. Hirsch: „Aurora-5 Experimental Framework for the Performance Evaluation of Speech Recognition in Case of a Hands-free Speech Input in Noisy Environments“, Tech. Rep., Niederrhein University of Applied Sciences, 2007.

[HM94] H. Hermansky und N. Morgan: „RASTA Processing of Speech“, *IEEE Transactions on Speech and Audio Processing*, Band 2(4), S. 578–589, Okt. 1994.

[HMBK91] H. Hermansky, N. Morgan, A. Bayya und P. Kohn: „The Challenge of Inverse-E: the RASTA-PLP Method“, *Proc. of Asilomar Conference on Signals, Systems and Computers (ACSSC)*, S. 800–804, Pacific Grove, CA, USA, Nov. 1991.

[HNKT00] S. Hirobayashi, H. Nomura, T. Koike und M. Tohyama: „Speech Waveform Recovery from a Reverberant Speech Signal Using Inverse Filtering of the Power Envelope Transfer Function“, *Electronics and Communications in Japan (Part III: Fundamental Electronic Science)*, Band 83(6), S. 77–85, 2000.

[HS85] T. Houtgast und H. J. M. Steeneken: „A Review of the MTF Concept in Room Acoustics and its Use for Estimating Speech Intelligibility in Auditoria“, *The Journal of the Acoustical Society of America*, Band 77(3), S. 1069–1077, 1985.

[HSP80] T. Houtgast, H. Steeneken und R. Plomp: „Predicting Speech Intelligibility in Rooms from the Modulation Transfer Function I General Room Acoustics“, *Acustica*, Band 46(1), S. 60–72, 1980.

[IFN10] O. Ichikawa, T. Fukuda und M. Nishimura: „Dynamic Features in the Linear-Logarithmic Hybrid Domain for Automatic Speech Recognition in a Reverberant Environment“, *IEEE Journal of Selected Topics in Signal Processing*, Band 4(5), S. 816–823, Okt. 2010.

[Int96] International Telecommunication Union (ITU), Geneva, Switzerland: *Recommendation G.712 – Transmission Performance Characteristics of Pulse Code Modulation Channels*, Nov. 1996.

[Iss18] L. Isserlis: „On a Formula for the Product-Moment Coefficient of Any Order of a Normal Frequency Distribution in Any Number of Variables“, *Biometrika*, Band 12(1-2), S. 134–139, 1918.

[KDNM09] K. Kinoshita, M. Delcroix, T. Nakatani und M. Miyoshi: „Suppression of Late Reverberation Effect on Speech Signal Using Long-Term Multiple-step Linear Prediction“, *IEEE Transactions on Audio, Speech, and Language Processing*, Band 17(4), S. 534–545, Mai 2009.

[KHU10] A. Krueger und R. Haeb-Umbach: „Model-Based Feature Enhancement for Reverberant Speech Recognition“, *IEEE Transactions on Audio, Speech, and Language Processing*, Band 18(7), S. 1692–1707, 2010.

[Kim94] C.-J. Kim: „Dynamic Linear Models with Markov-Switching“, *Journal of Econometrics*, Band 60(1-2), S. 1–22, 1994.

[KK09] K.-D. Kammeyer und K. Kroschel: *Digitale Signalverarbeitung - Filterung und Spektralanalyse mit MATLAB®-Übungen*, Vieweg+Teubner-Verlag, Wiesbaden, 7. Aufl., Apr. 2009.

[KLHU+10] A. Krueger, V. Leutnant, R. Haeb-Umbach, A. Marcel und J. Bloemer: „On the Initialization of Dynamic Models for Speech Features“, *Proc. of ITG Fachtagung Sprachkommunikation*, Bochum, Okt. 2010.

[KM97] B. Kingsbury und N. Morgan: „Recognizing Reverberant Speech with RASTA-PLP“, *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, S. 1259–1262, Munich, Germany, Apr. 1997.

[KMG98] B. E. D. Kingsbury, N. Morgan und S. Greenberg: „Robust Speech Recognition Using the Modulation Spectrogram“, *Speech Communication*, Band 25(1-3), S. 117–132, 1998.

[KNM05] K. Kinoshita, T. Nakatani und M. Miyoshi: „Fast Estimation of a Precise Dereverberation Filter based on Speech Harmonicity“, *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, S. 1073–1076, Philadelphia, PA, USA, 2005.

[KNM06] K. Kinoshita, T. Nakatani und M. Miyoshi: „Spectral Subtraction Steered by Multi-Step Forward Linear Prediction For Single Channel Speech Dereverberation“, *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, S. 817–820, Toulouse, France, 2006.

[Kut00] H. Kuttruff: *Room Acoustics*, Spon Press, London, UK, 4. Aufl., 2000.

[Kut04] H. Kuttruff: *Akustik: Eine Einführung*, S. Hirzel Verlag, 2004.

[LBD01] K. Lebart, J. Boucher und P. Denbigh: „A New Method Based on Spectral Subtraction for Speech Dereverberation“, *Acta Acustica united with Acustica*, Band 87, S. 359–366(8), 2001.

[LS82] T. Langhans und H. Strube: „Speech Enhancement by Nonlinear Multiband Envelope Filtering“, *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, S. 156–159, Paris, France, Mai 1982.

[Mak75] J. Makhoul: „Linear Prediction: A Tutorial Review“, *Proceedings of the IEEE*, Band 63(4), S. 561–580, Apr. 1975.

[Man64] J. Mandel: *The Statistical Analysis of Experimental Data*, Interscience, New York, 1964.

[MH83] J. Mourjopoulos und J. Hammond: „Modelling and Enhancement of Reverberant Speech Using an Envelope Convolution Method“, *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Band 8, S. 1144–1147, Boston, MA, USA, Apr. 1983.

[MK88] M. Miyoshi und Y. Kaneda: „Inverse Filtering of Room Acoustics“, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Band 36(2), S. 145–152, Febr. 1988.

[MM10] H. K. Maganti und M. Matassoni: „An Auditory Based Modulation Spectral Feature for Reverberant Speech Recognition“, *Proc. of Annual Conference of the International Speech Communication Association (Interspeech)*, S. 570–573, Makuhari, Japan, Sept. 2010.

[MOG00] M. Matassoni, M. Omologo und D. Giuliani: „Hands-free Speech Recognition Using a Filtered Clean Corpus and Incremental HMM Adaptation“, *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, S. 1407–1410, Istanbul, Turkey, Juni 2000.

[Mou85] J. Mourjopoulos: „On the Variation and Invertibility of Room Impulse Response Functions“, *Journal of Sound and Vibration*, Band 102(2), S. 217–228, 1985.

[MS95] J. Makhoul und R. Schwartz: „State of the Art in Continuous Speech Recognition“, *Proceedings of the National Academy of Sciences of the United States of America*, Band 92(22), S. 9956–9963, 1995.

[Mur98] K. Murphy: „Switching Kalman Filters“, Tech. Rep., U.C. Berkeley, 1998.

[NA79] S. T. Neely und J. B. Allen: „Invertibility of a Room Impulse Response“, *The Journal of the Acoustical Society of America*, Band 66(1), S. 165–169, 1979.

[NJKM05] T. Nakatani, B.-H. Juang, K. Kinoshita und M. Miyoshi: „Harmonicity Based Dereverberation with Maximum A Posteriori Estimation“, *Proc. of IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*, S. 94–97, New Paltz, NY, USA, Okt. 2005.

[NKM07] T. Nakatani, K. Kinoshita und M. Miyoshi: „Harmonicity-Based Blind Dereverberation for Single-Channel Speech Signals“, *IEEE Transactions on Audio, Speech, and Language Processing*, Band 15(1), S. 80–95, Jan. 2007.

[NM03] T. Nakatani und M. Miyoshi: „Blind Dereverberation of Single Channel Speech Signal Based on Harmonic Structure“, *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, S. 92–95, Hong Kong, Hong Kong, Apr. 2003.

[NMK05] T. Nakatani, M. Miyoshi und K. Kinoshita: „Single-Microphone Blind Dereverberation“, T. Nakatani, M. Miyoshi und K. Kinoshita, Hrsg., *Speech Enhancement*, Signals and Communication Technology, S. 247–270, Springer Berlin Heidelberg, 2005.

[NYK+08] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi und B.-H. Juang: „Blind Speech Dereverberation with Multi-Channel Linear Prediction Based on Short Time Fourier Transform Representation“, *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, S. 85–88, Las Vegas, NV, USA, Apr. 2008.

[OSB99] A. V. Oppenheim, R. W. Schafer und J. R. Buck: *Discrete-Time Signal Processing*, Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 2. Aufl., 1999.

[PB92] D. B. Paul und J. M. Baker: „The Design for the Wall Street Journal-based CSR corpus“, *Proc. of International Conference on Spoken Language Processing (ICSLP)*, S. 899–902, Banff, Alberta, Canada, Okt. 1992.

[PBB02] K. J. Palomaki, G. J. Brown und J. Barker: „Missing Data Speech Recognition in Reverberant Conditions“, *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, S. 65–68, Orlando, FL, USA, Mai 2002.

[PBB04] K. J. Palomäki, G. J. Brown und J. P. Barker: „Techniques for Handling Convolutional Distortion with ‘Missing Data’ Automatic Speech Recognition“, *Speech Communication*, Band 43(1-2), S. 123–142, 2004.

[PLLH08] R. Petrick, K. Lohde, M. Lorenz und R. Hoffmann: „A New Feature Analysis Method for Robust ASR in Reverberant Environments Based on the Harmonic Structure of Speech“, *Proc. of European Signal Processing Conference (EUSIPCO)*, Lausanne, Switzerland, Aug. 2008.

[PLU+08] R. Petrick, X. Lu, M. Unoki, M. Akagi und R. Hoffmann: „Robust Front End Processing for Speech Recognition in Reverberant Environments: Utilization of Speech Characteristics“, *Proc. of Annual Conference of the International Speech Communication Association (Interspeech)*, S. 658–661, Brisbane, Australia, Sept. 2008.

[Pol88] J. Polack: *La Transmission de l’Énergie Sonore dans les Salles*, Dissertation, Université du Maine, 1988.

[PP02] N. Parihar und J. Picone: „DSR Front End LVCSR Evaluation“, Tech. Rep. AU/384/02, Aurora Working Group, 2002.

[PP08] K. B. Petersen und M. S. Pedersen: *The Matrix Cookbook*, Technical University of Denmark, Okt. 2008, [URL] http://www2.imm.dtu.dk/pubdb/p.php?3274, Version 20081110.

[PRH+92] R. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang und M. Allerhand: „Complex Sounds and Auditory Images“, Y. Cazals, L. Demany, und K. Horner, Hrsg., *Auditory Physiology and Perception*, S. 429–446, Pergamon, Oxford, 1992.

[PS06] F. Pacheco und R. Seara: „Spectral Subtraction for Reverberation Reduction Applied to Automatic Speech Recognition“, *Proc. of International Telecommunications Symposium (ITS)*, S. 795–800, Fortaleza, Ceara, Brazil, Sept. 2006.

[QC93] S. Qian und D. Chen: „Discrete Gabor Transform“, *IEEE Transactions on Signal Processing*, Band 41(7), S. 2429–2438, Juli 1993.

[Ric09] P. Rico: *Robuste Spracherkennung unter raumakustischen Umgebungsbedingungen*, Dissertation, Technische Universität Dresden, 2009.

[RJ93] L. Rabiner und B. H. Juang: *Fundamentals of Speech Recognition*, Prentice Hall, 1993.

[RJO04] R. Ratnam, D. Jones und J. O'Brien, W.D.: „Fast Algorithms for Blind Estimation of Reverberation Time", *IEEE Signal Processing Letters*, Band 11(6), S. 537 – 540, Juni 2004.

[RJW+03] R. Ratnam, D. L. Jones, B. C. Wheeler, W. D. O'Brien, C. R. Lansing und A. S. Feng: „Blind Estimation of Reverberation Time", *The Journal of the Acoustical Society of America*, Band 114(5), S. 2877–2892, Nov. 2003.

[RLS94] A. E. Rosenberg, C.-H. Lee und F. K. Soong: „Cepstral Channel Normalization Techniques for HMM-Based Speaker Verification", *Proc. of International Conference on Spoken Language Processing (ICSLP)*, S. 1835–1838, 1994.

[RNS05a] C. Raut, T. Nishimoto und S. Sagayama: „Maximum Likelihood Based HMM State Filtering Approach to Model Adaptation for Long Reverberation", *Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, S. 353–356, Nov. 2005.

[RNS05b] C. K. Raut, T. Nishimoto und S. Sagayama: „Acoustic Model Adaptation for Reverberant Speech by State Splitting of HMM and Convolution of Distributions", *Techn. Report of Institute of Electronics, Information and Communication Engineers (IEIC)*, Band 104, S. 37–42, 2005.

[RNS05c] C. K. Raut, T. Nishimoto und S. Sagayama: „Model Adaptation by State Splitting of HMM for Long Reverberation", *Proc. of European Conference on Speech Communication and Technology (EUROSPEECH)*, S. 277–280, Lisbon, Portugal, Sept. 2005.

[RWK00] B. Radlovic, R. Williamson und R. Kennedy: „Equalization in an Acoustic Reverberant Environment: Robustness Results", *IEEE Transactions on Speech and Audio Processing*, Band 8(3), S. 311–319, Mai 2000.

[SC00] M. Shire und B. Chen: „Data-Driven RASTA Filters in Reverberation", *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, S. 1627–1630, Istanbul, Turkey, Juni 2000.

[Sch65] M. R. Schroeder: „New Method of Measuring Reverberation Time", *The Journal of the Acoustical Society of America*, Band 37(6), S. 1187–1188, 1965.

[SCI75] J. Stockham, T.G., T. Cannon und R. Ingebretsen: „Blind Deconvolution Through Digital Signal Processing", *Proceedings of the IEEE*, Band 63(4), S. 678–692, Apr. 1975.

[SFB01] V. Stahl, A. Fischer und R. Bippus: „Acoustic Synthesis of Training Data for Speech Recognition in Living Room Environments“, *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, S. 21–24, Salt Lake City, Utah, Mai 2001.

[SK08] A. Sehr und W. Kellermann: „Towards Robust Distant-Talking Automatic Speech Recognition in Reverberant Environments“, E. Hänsler und G. Schmidt, Hrsg., *Speech and Audio Processing in Adverse Environments*, Signals and Communication Technology, S. 679–728, Springer Berlin Heidelberg, 2008.

[SMK10] A. Sehr, R. Maas und W. Kellermann: „Reverberation Model-Based Decoding in the Logmelspec Domain for Robust Distant-Talking Speech Recognition“, *IEEE Transactions on Audio, Speech, and Language Processing*, Band 18(7), S. 1676–1691, 2010.

[SMK11] A. Sehr, R. Maas und W. Kellermann: „Frame-Wise HMM Adaptation Using State-Dependent Reverberation Estimates“, *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republik, Mai 2011.

[SPW96] S. Subramaniam, A. Petropulu und C. Wendt: „Cepstrum-Based Deconvolution for Speech Dereverberation“, *IEEE Transactions on Speech and Audio Processing*, Band 4(5), S. 392–396, Sept. 1996.

[ST95] E. G. Schukat-Talamazzini: *Automatische Spracherkennung - Grundlagen, statistische Modelle und effiziente Algorithmen*, Künstliche Intelligenz, Vieweg, 1995.

[SZK06] A. Sehr, M. Zeller und W. Kellermann: „Distant-Talking Continuous Speech Recognition Based on a Novel Reverberation Model in the Feature Domain“, *Proc. of Annual Conference of the International Speech Communication Association (Interspeech)*, S. 769–772, Pittsburgh, PA, USA, Sept. 2006.

[TGH08a] S. Thomas, S. Ganapathy und H. Hermansky: „Hilbert Envelope Based Features for Far-Field Speech Recognition“, *Proc. of Joint Workshop on Machine Learning and Multimodal Interaction (MLMI)*, S. 119–124, Utrecht, The Netherlands, Sept. 2008.

[TGH08b] S. Thomas, S. Ganapathy und H. Hermansky: „Recognition of Reverberant Speech Using Frequency Domain Linear Prediction“, *IEEE Signal Processing Letters*, Band 15, S. 681–684, 2008.

[TN04] T. Takiguchi und M. Nishimura: „Acoustic Model Adaptation Using First Order Prediction for Reverberant Speech“, *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, S. 869–872, Montreal, Quebec, Canada, Mai 2004.

[TS05] M. Triki und D. T. M. Slock: „Blind Dereverberation of a Single Source Based on Multichannel Linear Prediction“, *Proc. of International Workshop on Acoustic Echo and Noise Control (IWAENC)*, S. 173–176, Eindhoven, The Netherlands, Sept. 2005.

[TTN06] A. M. Toh, R. Togneri und S. Nordholm: „Combining MLLR Adaptation and Feature Extraction for Robust Speech Recognition in Reverberant Environments“, *Proc. of International Conference on Speech Science and Technology (SST)*, S. 88–93, Auckland, New Zealand, Dez. 2006.

[TTN07] A. M. Toh, R. Togneri und S. Nordholm: „Feature and Distribution Normalization Schemes for Statistical Mismatch Reduction in Reverberant Speech Recognition“, *Proc. of Annual Conference of the International Speech Communication Association (Interspeech)*, S. 234–237, Antwerp, Belgium, Aug. 2007.

[TW02] F. Talantzis und D. Ward: „Investigation of Performance of Acoustic Arrays for Equalization in a Reverberant Environment“, *Proc. of International Conference on Digital Signal Processing (DSP)*, S. 247–250, Santorini, Greece, Juli 2002.

[UFSA03] M. Unoki, M. Furukawa, K. Sakata und M. Akagi: „A Method Based on the MTF Concept for Dereverberating the Power Envelope from the Reverberant Signal“, *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, S. 888–891, Hong Kong, China, Apr. 2003.

[UN98] N. Ueda und R. Nakano: „Deterministic Annealing EM Algorithm“, *Neural Networks*, Band 11(2), S. 271–282, 1998.

[VL98] O. Viikki und K. Laurila: „Cepstral Domain Segmental Feature Vector Normalization for Noise Robust Speech Recognition“, *Speech Communication*, Band 25(1-3), S. 133–147, 1998.

[VM06] P. Vary und R. Martin: *Digital Speech Transmission: Enhancement, Coding and Error Concealment*, JohnWiley & Sons, 2006.

[vVH97] S. van Vuuren und H. Hermansky: „Data-Driven Design of RASTA-Like Filters“, *Proc. of European Conference on Speech Communication and Technology (EUROSPEECH)*, S. 409–412, Rhodes, Greece, Sept. 1997.

[WHN08] J. Wen, E. Habets und P. Naylor: „Blind Estimation of Reverberation Time Based on the Distribution of Signal Decay Rates“, *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, S. 329 – 332, Las Vegas, USA, Apr. 2008.

[Wöl09] M. Wölfel: „Enhanced Speech Features by Single-Channel Joint Compensation of Noise and Reverberation“, *IEEE Transactions on Audio, Speech, and Language Processing*, Band 17(2), S. 312–323, Febr. 2009.

[WR90] J. Wexler und S. Raz: „Discrete Gabor Expansions“, *Signal Processing*, Band 21(3), S. 207–220, Nov. 1990.

[WSNK09] J. Y. C. Wen, A. Sehr, P. A. Naylor und W. Kellermann: „Blind Estimation of a Feature-Domain Reverberation Model in Non-Diffuse Environments with Variance Adjustment“, *Proc. of European Signal Processing Conference (EUSIPCO)*, S. 175–178, Glasgow, Scotland, Aug. 2009.

[YEG+06] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev und P. C. Woodland: *The HTK Book, version 3.4*, Cambridge University Engineering Department, Cambridge, UK, 2006.

[YM00] B. Yegnanarayana und P. Murthy: „Enhancement of Reverberant Speech Using LP Residual Signal“, *IEEE Transactions on Speech and Audio Processing*, Band 8(3), S. 267–281, Mai 2000.

[YNM09] T. Yoshioka, T. Nakatani und M. Miyoshi: „Integrated Speech Enhancement Method Using Noise Suppression and Dereverberation“, *IEEE Transactions on Audio, Speech, and Language Processing*, Band 17(2), S. 231–246, Febr. 2009.

[YNS04] H. Yamamoto, T. Nishimoto und S. Sagayama: „Frame-by-Frame HMM Adaptation for Reverberant Speech Recognition“, *Proc. of Special Workshop in Maui (SWIM)*, Maui, Jan. 2004.

[You08] S. Young: „HMMs and Related Speech Recognition Technologies“, J. Benesty, M. Mohan Sondhi und Y. Huang, Hrsg., *Springer Handbook of Speech Processing*, Kap. 27, Springer, Berlin, 2008.

Eigene Publikationen

[HUKus] R. Haeb-Umbach und A. Krueger: „Reverberant Speech Recognition“, T. Virtanen, B. Raj und R. Singh, Hrsg., *Techniques for Noise Robustness in Automatic Speech Recognition*, Kap. 10, John Wiley & Sons, Ltd., Veröffentlichung steht noch aus.

[KHU09] A. Krueger und R. Haeb-Umbach: „Model Based Feature Enhancement for Automatic Speech Recognition in Reverberant Environments“, *Proc. of Annual Conference of the International Speech Communication Association (Interspeech)*, S. 1231–1234, Brighton, U.K., Sept. 2009.

[KHU10] A. Krueger und R. Haeb-Umbach: „Model-Based Feature Enhancement for Reverberant Speech Recognition“, *IEEE Transactions on Audio, Speech, and Language Processing*, Band 18(7), S. 1692–1707, 2010.

[KHU11a] A. Krueger und R. Haeb-Umbach: „MAP-Based Estimation of the Parameters of Non-Stationary Gaussian Processes from Noisy Observations“, *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, Mai 2011.

[KHU11b] A. Krueger und R. Haeb-Umbach: „A Model Based Approach to Joint Compensation of Noise and Reverberation for Speech Recognition“, R. Haeb-Umbach und D. Kolossa, Hrsg., *Robust Speech Recognition of Uncertain or Missing Data*, Kap. 10, Springer, 2011.

[KLHU$^+$10] A. Krueger, V. Leutnant, R. Haeb-Umbach, A. Marcel und J. Bloemer: „On the Initialization of Dynamic Models for Speech Features“, *Proc. of ITG Fachtagung Sprachkommunikation*, Bochum, Okt. 2010.

[KWHU08] A. Krueger, E. Warsitz und R. Haeb-Umbach: „Blinde Akustische Strahlformung für Anwendungen im KFZ“, *Proc. of Deutsche Jahrestagung für Akustik (DAGA)*, S. 863–864, Dresden, März 2008.

[KWHU11] A. Krueger, E. Warsitz und R. Haeb-Umbach: „Speech Enhancement With a GSC-Like Structure Employing Eigenvector-Based Transfer Function Ratios Estimation“, *IEEE Transactions on Audio, Speech, and Language Processing*, Band 19(1), S. 206–219, 2011.

[LKHU11] V. Leutnant, A. Krueger und R. Haeb-Umbach: „A Versatile Gaussian Splitting Approach to Non-Linear State Estimation and Its Application to Noise-Robust ASR“, *Proc. of Annual Conference of the International Speech Communication Association (Interspeech)*, Florence, Italy, Aug. 2011.

[RWKHU10] B. Raj, K. Wilson, A. Krueger und R. Haeb-Umbach: „Ungrounded Independent Non-Negative Factor Analysis“, *Proc. of Annual Conference of the International Speech Communication Association (Interspeech)*, S. 330–333, Makuhari, Japan, Sept. 2010.

[TVKHU08] D. H. Tran Vu, A. Krueger und R. Haeb-Umbach: „Generalized Eigenvector Blind Speech Separation Under Coherent Noise in a GSC Configuration“, *Proc. of International Workshop on Acoustic Echo and Noise Control (IWAENC)*, Seattle, Washington, USA, Sept. 2008.

[WKHU08] E. Warsitz, A. Krueger und R. Haeb-Umbach: „Speech Enhancement With a New Generalized Eigenvector Blocking Matrix for Application in a Generalized Sidelobe Canceller“, *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, S. 73–76, Las Vegas, NV, USA, März 2008.