UNIVERSITÄT PADERBORN
Die Universität der Informationsgesellschaft

FAKULTÄT FÜR
ELEKTROTECHNIK,
INFORMATIK und
MATHEMATIK

Beiträge zur generalisierten modellbasierten spektralen Sprachsignalentstörung

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.-Ing. Aleksej Chinaev

Erster Gutachter:	Prof. Dr.-Ing. Reinhold Häb-Umbach
Zweiter Gutachter:	Prof. Dr.-Ing. Rainer Martin
Tag der mündlichen Prüfung:	01.12.2017

Paderborn 2018

Diss. EIM-E/338

Danksagung

Während meiner etwas mehr als siebenjährigen Beschäftigung als Wissenschaftlicher Mitarbeiter im Fachgebiet Nachrichtentechnik der Universität Paderborn verfasste ich die vorliegende Arbeit. Dabei begleiteten und unterstützten mich viele wunderbare Menschen, bei denen ich mich an dieser Stelle herzlich bedanken möchte.

Zunächst möchte ich mich jedoch bei meinem himmlischen Vater dafür bedanken, dass er mich so schuf, wie ich bin, und bei meinem Herrn Jesus Christus dafür, dass er mir Gottes grenzenlose Liebe offenbarte. Ich danke auch meinen Eltern und Großeltern dafür, dass sie mich in Liebe aufzogen und zu akademischer Laufbahn ermutigten. Mein besonderer Dank gilt meiner Mutter, Elfriede Pauls, für ihre große Liebe zu mir. Auch meiner geliebten Ehefrau, Irina, die mir immer treu zur Seite steht, möchte ich aufs Herzlichste danken. Ich danke meinen lieben Kindern, Evnika*, Efim und Karina, für ihre ansteckende Lebensfreude. Auch bei allen Geschwistern der "Christus für alle" Gemeinde in Paderborn und in Bielefeld möchte ich mich für die geistliche Unterstützung und die vielen Gebete bedanken.

Besonders möchte ich mich bei Herrn Prof. Dr.-Ing. Reinhold Häb-Umbach bedanken, für die Möglichkeit unter seiner akademischen Vaterschaft zu promovieren. Danke für das entgegengebrachte Vertrauen und für die kompetente Begleitung. Ich danke auch allen meinen ehemaligen und jetzigen Kollegen für das angenehme Arbeitsklima, die persönlichen Freundschaften, die fruchtbaren Gespräche, die kompetenten Anregungen, die hilfreichen softwaretechnischen Tipps, die geteilten Softwarekomponenten und die konstruktiven Anmerkungen zu dieser Arbeit. Danke euch, ihr lieben Ehemaligen, Sven Peschke, Maik Bevermeier, Alexander Krüger, Dang Hai Tran Vu, Volker Leutnant, Manh Kha Hoang, Florian Jacob, Oliver Walter*, Prerna Arora*, Vladimir Despotovic, Ursula Stiebritz und Christine Fricke[1]. Danke euch, ihr lieben Jetzigen, Jörg Schmalenströer, Jörg Ullmann, Peter Schütte, Lukas Drude*, Jahn Heymann, Thomas Glarner, Jens Heitkämper, Jannek Ebbers und Anita Hüser. Außerdem gilt mein Dank den Studenten, Robin Moritz, Stefan Beller, Johann Veer, Paul Traut, Mark Püls, Maximilian Schreckenberg, Patrick Hanebrink, Eduard Schwab und Robin Wolf*, deren Abschlussarbeiten ich betreuen durfte. Noch danke ich Herrn Prof. Dr.-Ing. Rainer Martin für Übernahme des Korreferats dieser Arbeit.

Ansonsten danke ich allen, die in Teilen diese Arbeit zur Korrektur aufmerksam lasen[2].

"Und alles, was ihr tut mit Worten oder mit Werken, das tut alles im Namen des Herrn Jesus und dankt Gott, dem Vater, durch ihn."

Apostel Paulus, Kolosser 3:17 [Bib84]

[1] Aufgrund des Platzmangels wird auf Angabe akademischer Titel verzichtet.

[2] Entsprechende Namen werden im Text mit einem Sternchen als kleine Auszeichnung markiert.

UNIVERSITÄT PADERBORN
Die Universität der Informationsgesellschaft

FAKULTÄT FÜR
ELEKTROTECHNIK,
INFORMATIK und
MATHEMATIK

Abstract

The ongoing digitalization of our society has led to new requirements on digital signal processing concerning performance, robustness, and efficiency. Based on the physical nature of speech, speech signals are often processed in the time-frequency domain, where the psychoacoustic and statistical models can be more easily integrated into the signal processing chain as compared to the time domain. An important processing task invariably occurring in many technical fields is denoising of signals disturbed by additive noise. A conventional spectral speech enhancement system for signal denoising usually consists of the following modules:

- Module 1: Estimation of the power spectral density (PSD) of the interfering signal,
- Module 2: Calculation of the spectral gain function for the speech denoising,
- Module 3: Estimation of the so-called *a priori* signal-to-noise ratio (SNR),
- Module 4: Calculation of the speech presence probability (SPP).

While stationary noise can be efficiently suppressed by such systems, removal of nonstationary noise has been known to be a very demanding task, whose solution is still a challenging topic of modern research. Development of modules being able to remove nonstationary noise, are carried out in two main parts of this thesis.

The first part addresses development of three approaches, that can be used as Module 1 for noise PSD tracking: a *Minimum Statistics* approach with a new Bayesian-motivated control function, a noise tracker based on the noise-only presence probability calculated using a deep neural network (DNN), and a *maximum a posteriori* based approach aimed at improving the noise PSD estimates of any conventional noise tracker, to which it acts as a postprocessor. While the first method aims at increasing the intelligibility of processed signals, the goal of the other two approaches is to improve the signal quality while suppressing the noise signal as much as possible. Experimental evaluation shows that as compared to the state-of-the-art methods, the DNN-based noise PSD tracker achieves the best performance and leads to both the best signal quality and the strongest noise reduction.

In the second part of this work, three methods are presented, which can act as Modules 2, 3 and 4, respectively: a spectral gain function of generalized log-spectral amplitudes, a generalized *Decision-Directed* approach for *a priori* SNR estimation and a SPP estimator based on weighted generalized *a posteriori* SNR estimates. All these approaches are designed to work in domains of the so-called generalized spectral values, which are introduced as logarithmic generalized spectral amplitudes and as generalized SNR values. Compared to the state-of-the-art approaches, signal processing in the generalized domains leads to stronger noise suppression with slightly improved signal quality.

UNIVERSITÄT PADERBORN
Die Universität der Informationsgesellschaft

FAKULTÄT FÜR
ELEKTROTECHNIK,
INFORMATIK und
MATHEMATIK

Kurzfassung

Mit zunehmender Digitalisierung unserer Gesellschaft wachsen auch Anforderungen an die Systeme zur Verarbeitung von zeitdiskreten Sprachsignalen bezüglich ihrer Leistungsfähigkeit, Robustheit und Effizienz. Begründet in der physikalischen Natur eines Sprachsignals findet die Sprachsignalverarbeitung oft im Zeit-Frequenz-Bereich statt, wo die psychoakustischen und statistischen Modelle in die Signalverarbeitungskette leichter integriert werden können als im Zeitbereich. Häufig müssen hier die gestörten Sprachsignale vom additiven Rauschen befreit werden, das als eine prominente Störung in vielen technischen Bereichen auftritt. Dabei werden Systeme zur spektralen Sprachsignalentstörung modular aus verschiedenen Systemkomponenten aufgebaut, welche die folgenden Aufgaben übernehmen:

- Baustein 1: Schätzung des Rauschleistungsdichtespektrums (RLDS),
- Baustein 2: Berechnung einer Filterfunktion zur Entstörung spektraler Amplituden,
- Baustein 3: Schätzung des sogenannten *a priori* SNR,
- Baustein 4: Berechnung der Wahrscheinlichkeit für Sprachsignalpräsenz.

Während solche Systeme stationäres Rauschen gut unterdrücken können, stellt sich das spektrale Entfernen nichtstationärer Störungen als eine sehr herausfordernde Aufgabe dar, die immer noch Gegenstand moderner Forschung ist. In zwei Hauptteilen der vorliegenden Arbeit werden mit unterschiedlichen Zielsetzungen sechs Schätzverfahren entwickelt, die als Bausteine eines Systems zur spektralen Sprachsignalentstörung verwendet werden.

Der Hauptaugenmerk des ersten Teils liegt dabei auf der Entwicklung von drei Verfahren zur RLDS-Schätzung, die jeweils als Baustein 1 eingesetzt werden: das *Minimum Statistics* Verfahren mit einer neuen Bayes-motivierten Steuerungsfunktion, ein Schätzer basierend auf den spektralen Masken, die mit einem neuronalen Netz berechnet werden, und ein *maximum a posteriori* Schätzer, der als Postprozessor zu einem beliebigen Rauschschätzer agieren kann. Während das erste Verfahren dazu dient, die Verständlichkeit gestörter Sprachsignale zu erhöhen, verfolgen die beiden anderen Verfahren das Ziel, Sprachsignalqualität bei einer möglichst hohen Störsignaldämpfung zu verbessern.

Im zweiten Teil dieser Arbeit werden drei Verfahren entwickelt, die im Bereich der so genannten generalisierten spektralen Größen arbeiten und jeweils als Baustein 2, 3 und 4 eingesetzt werden: eine spektrale Filterfunktion der generalisierten logarithmischen spektralen Amplituden, ein generalisiertes *Decision-Directed* Verfahren zur Schätzung des *a priori* Signal-zu-Rausch Verhältnisses (engl. *signal-to-noise ratio*, SNR) und ein Schätzer der Sprachpräsenzwahrscheinlichkeit, der auf den gewichteten generalisierten *a posteriori* SNR-Werten arbeitet. Neu für die Sprachsignalverarbeitung ist hier die Einführung der unkonventionellen generalisierten Größen, die zur besseren Sprachsignalentstörung führen.

Inhaltsverzeichnis

1. Einleitung

Gesprochene Sprache begleitet uns durchs Leben und dient als zentrales Mittel zwischenmenschlicher Verständigung. Sie ist ein akustisches Phänomen und eilt der geschriebenen Sprache voraus, denn in der Regel lernt ein Mensch zunächst sprechen und dann schreiben. Ohne mündliche Kommunikation sind frühkindliche Erziehung, fundierte Allgemeinbildung oder funktionierende zwischenmenschliche Beziehungen nur schwer vorstellbar. Das ist die menschliche Stimme, die den Worten den emotionalen Rahmen und die einmalige Individualität verleiht. Von den Gefühlen moduliert, offenbart sich die menschliche Stimme als Spiegelbild unserer Seele [BBB+13]. Das ausgesprochene Wort formt, beeinflusst und berührt, denn es ist nicht nur wichtig, was gesagt wird, sondern auch wie. Glaubt man der Bibel, so wurde das ganze Universum durch das gesprochene Wort Gottes erschaffen [Bib84]. Was für eine geistliche Autorität kann das ausgesprochene Wort haben!

Viele Jahrhunderte dienten Schallwellen als einziger Informationsträger der gesprochenen Sprache und ein Sprachsignal war lange Zeit nur ein Schallsignal. Hohe Luftabsorption der Schallenergie begrenzt jedoch den Wirkungsbereich von Sprachsignalen, sodass gewisse räumliche Nähe der Gesprächspartner eine notwendige Bedingung für Sprachkommunikation war. Bestrebungen der Menschen, diese räumliche Barrieren zu überwinden, wurden erst am Ende des 19. Jahrhunderts mit der Entwicklung der drahtgebundenen Telefonie und später des drahtlosen Hörfunks belohnt. Sie ermöglichten zum ersten Mal die Sprachkommunikation weit über die Grenzen der Schallausbreitung hinaus [Mar90]. Zu weiteren Informationsträgern der gesprochenen Sprache werden dabei die elektrischen Signale in den elektrischen Leitungen und die elektromagnetischen Wellen in der Luft. Mit Hilfe der analogen Technik können jetzt Sprachsignale sowohl aufgenommen, übertragen und wiedergegeben als auch gespeichert und verarbeitet werden. Somit beginnt die Ära der analogen maschinellen Sprachsignalverarbeitung.

Die Digitalisierung der Technik in der zweiten Hälfte des 20. Jahrhunderts führte zur Entwicklung innovativer Technologien der modernen Sprachkommunikation [RS07]. Rasante Fortschritte in der Herstellung leistungsfähiger Endgeräte mit kostengünstiger Mikroelektronik und die damit verbundene schnelle Verbreitung vom Mobilfunk und Internet trugen zur weiteren Popularisierung der drahtlosen Kommunikationstechnik in der gesamten Bevölkerung bei. Die weltweite Sprachkommunikation wurde damit mobil und kostengünstig. Gesprochene Sprache bekommt dabei einen weiteren Informationsträger die zeitdiskreten digitalen Sprachsignale, die durch die zeitliche Abtastung und Amplitudenquantisierung der elektrischen Sprachsignale gewonnen werden. Die Möglichkeit, die zeitdiskreten Signale in Signalprozessoren zu verarbeiten, eröffnet für die moderne Sprachsignalverarbeitung neue bisher ungeahnte Horizonte. Die digitale Sprachsignalverarbeitung gewinnt somit an Bedeutung in der Kommunikationstechnik und wird zu einem eigenständigen Teilgebiet der modernen Informationstechnik angesiedelt in den Ingenieurwissenschaften [VHH98].

Die digitale Sprachsignalverarbeitung ist vielfältig und umfasst verschiedene Bereiche, unter denen sich effiziente Sprachkodierung, akustische Sprachsynthese, automatische Spracherkennung und maschinelle Sprachsignalverbesserung etablierten [RS07]. Außerdem sind Sprechererkennung/Sprecherverifikation, akustische Szenenanalyse und auditive Analyse der Sprechermerkmale zu erwähnen[1]. Die Verfahren dieser Bereiche werden in den verschiedenen akustischen Systemen der Mensch-zu-Mensch oder Mensch-zu-Maschine Kommunikation eingesetzt. Da wir in einer geräuschvollen Umwelt leben, werden Verfahren zur Sprachsignalverbesserung gerne als eine Vorverarbeitungsstufe in vielen Anwendungen benutzt, in denen Eingangssignale nicht in Form von ungestörten Sprachsignalen vorliegen. Dies ist z. B. bei der robusten automatischen Spracherkennung der Fall [Loi13]. Während sie unter kontrollierbaren Bedingungen auf ungestörten Sprachsignalen ganz gut funktioniert, sinkt ihre Leistungsfähigkeit dramatisch für gestörte Sprachsignale. Weitere Anwendungen, in denen Verfahren zur Sprachsignalverbesserung eingesetzt werden, sind interaktive Telekonferenzsysteme, Freisprecheinrichtungen am Ohr oder im Auto mit integrierter Störunterdrückung, bidirektionale Sprachdialogsysteme, Kopfhörer mit aktiver Lärmkompensation, Hörgeräte für ältere Menschen oder für Hörgeschädigte, Wiederherstellung von archivierten Audioaufnahmen, Systeme zur Internet-Telefonie und nicht zuletzt Mobiltelefonie u. s. w. [BMC05, Vas08].

Grundsätzlich unterscheidet man dabei zwischen einer einkanaligen und einer mehrkanaligen Sprachsignalverbesserung. Bei der letzteren können Verfahren zur akustischen Strahlformung (engl. *beamforming*) eingesetzt werden, welche sich die räumliche Diversität verschiedener akustischer Quellen zunutze machen. Außerdem können Signale einer Mikrofongruppe auch zur Positionsbestimmung eines Sprechers, zur automatischen Sprecherverfolgung und sogar zur Kalibrierung der Mikrofongruppe verwendet werden, von der Sprachsignale aufgenommen werden [PJHUF16]. Für die Sprachsignalverbesserung eingesetzt, ist die akustische Strahlformung jedoch nicht immer imstande, Störsignale zu unterdrücken, die auf eine Mikrofongruppe aus der Richtung der Zielquelle einfallen. Aus diesem Grund wird einer mehrkanaligen Sprachsignalverbesserung in der Regel eine einkanalige Sprachsignalentstörung nachgeschaltet, die der Gegenstand dieser Arbeit ist.

Abhängig davon, in welchem Bereich die Signalverarbeitung stattfindet, unterscheidet man außerdem zwischen Zeitbereichs- oder Frequenzbereichsverfahren. Das wichtigste Argument für Sprachsignalverarbeitung im Frequenzbereich ist die Tatsache, dass ein Sprachsignal sich bereits bei seiner Erzeugung mit menschlichen Sprechorganen als Überlagerung von Schwingungen unterschiedlicher Frequenzen darstellt [BCH11b]. Für die Signalverarbeitung im Frequenzbereich spricht zum einen die bessere Trennung des Sprachsignals vom Störsignal, welche in der Dünnbesetztheit der Sprachsignalspektren begründet ist, zum anderen bessere statistische Dekorrelation der einzelnen spektralen Signalkomponenten, die aus diesem Grund unabhängig voneinander verarbeitet werden, und zum dritten die Möglichkeit, psychoakustische Modelle der Sprachsignale in die Signalverarbeitungskette zu integrieren [Mar03]. Diese Vorteile sorgen dafür, dass die Verarbeitung der Sprachsignale im Frequenzbereich bei vielen Anwendungen zum Einsatz kommt. Die Frage ist allerdings, welche Transformation für die Frequenzbereichsdarstellung am besten geeignet ist.

[1] Während die akustische Szenenanalyse Aufschluss über anwesende Personen und Gegebenheiten der Umgebung geben soll [Sch10], bezweckt auditive Analyse der Sprechermerkmale Erfassung solcher Sprechereigenschaften wie Emotionen, Alter und Geschlecht. Letztere kann zu Diagnosezwecken eingesetzt werden, ob der Sprecher unter einer Sprechstörung leidet oder momentan schläfrig oder betrunken ist [SSB+15].

Zur Auswahl stehen verschiedene Transformationen, wie die Kurzzeit-Fourier-Transformation (engl. *short-time Fourier transform*, STFT), die diskrete Kosinustransformation oder die diskrete Wavelet-Transformation [ADHG04]. Da die diskrete Kosinustransformation im Vergleich zur STFT eine etwas höhere Konzentration der Signalenergie und bessere Frequenzauflösung bei gleicher Fensterlänge aufweist, kann sie für die Sprachsignalverbesserung eingesetzt werden [SKY98]. Allerdings führt sie in ihrer konventionellen Definition zur sogenannten zeitlichen Verschiebungsvarianz und somit zu den unerwünschten Oszillationen ihrer Koeffizientenverläufe, was der Verwendung von reellwertigen Basisfunktionen geschuldet ist. Solche negativen Eigenschaften können Verluste der Leistungsfähigkeit der Verfahren zur Signalverarbeitung verursachen, welche die Korrelationen der aufeinander folgenden Koeffizienten von Sprachsignalen ausnutzen [DS09]. Die Wavelet-Transformation trägt zur noch besseren Auflösung der Zeit-Frequenz (ZF) Darstellung von Sprachsignalen bei, denn im Unterschied zu STFT stellt sie die niedrigen Frequenzanteile eines Signals frequenzscharf und die höheren Frequenzanteile zeitlich scharf dar [Wei09]. Außerdem existieren Wavelet-Transformationen mit komplexwertigen Basisfunktionen, mit deren Hilfe die gerade erwähnten Nachteile von reellwertigen Basisfunktionen vermieden werden. Allerdings hat sich die Wavelet-Transformation in der Sprachsignalverarbeitung im Unterschied zur Bildverarbeitung noch nicht so stark etabliert, wie dies bei der STFT der Fall ist [FBR04]. Die Popularität der STFT ist nicht zuletzt ihrer Einfachheit und ihrer effizienten Realisierung mittels der schnellen Fourier-Transformation (engl. *fast Fourier transform*, FFT) geschuldet [CR83]. Im Rahmen dieser Arbeit werden Verfahren zur einkanaligen Sprachsignalverbesserung entwickelt, welche im STFT-Bereich eingesetzt werden.

Einkanalige Sprachsignalentstörung im Frequenzbereich ist seit etwa 50 Jahren ein anerkanntes Forschungsgebiet [BMC05]. Und obwohl man sich schon relativ lange mit diesem Thema beschäftigt, gehört es immer noch zu den aktuellen Forschungsthemen und wird als *eine der schwierigsten Aufgaben in der Sprachsignalverbesserung* bezeichnet [Loi13]. Und der Schwierigkeitsgrad dieser Aufgabe wird klar, wenn man sich die große Vielfalt möglicher Sprachsignalstörungen vor Augen führt. So kann eine Störung mit dem Nutzsignal entweder unkorreliert oder korreliert sein. Im letzten Fall spricht man auch von einer konvolutiven Störung, wenn Sprachsignale z. B. in einem Raum aufgenommen werden. Außerdem kann eine Störung entweder omnipresent oder kurzzeitig sein. Kurzzeitige oder transiente akustische Störungen sind beispielsweise Geklapper auf einer Tastatur, Schläge eines Metronoms, Klopfen an der Tür. Besondere Aufmerksamkeit verdienen additive Störungen, die in der Sprachsignalverarbeitung sehr viele unterschiedliche Ausprägungen haben, denn das Nutzsignal wird vom Rauschen an verschiedenen Stellen eines Systems zur Sprachsignalverarbeitung additiv überlagert [Dav02]. Solche Störquellen wie Umgebungsrauschen, Sprachsignale der konkurrierenden Sprecher, Mikrofonrauschen, Verstärkerrauschen, Quantisierungsrauschen des Analog-Digital-Umsetzers aber auch das Rundungsrauschen des Digital-Analog-Umsetzers und Verluste von Audiocodec-Verfahren können durch das additive Rauschen modelliert werden. Ist die Störung konvolutiv wie bei Hall, verwendet man Verfahren zur akustischen Enthallung der Sprachsignale. Ist sie additiv, kommen Algorithmen zur Rauschunterdrückung zum Einsatz. Die Hauptaufgabe dieser Verfahren besteht darin, STFT-Koeffizienten des ungestörten Sprachsignals aus den STFT-Koeffizienten einer gestörten Sprachsignalaufnahme mittels einer Filterung zu gewinnen und zwar so, dass das Sprachsignal dabei möglichst wenig beschädigt wird [DO03]. Dies ist herausfordernd, denn in der Regel werden vom Störsignal Frequenzen belegt, die im Sprachsignal vorhanden sind.

Frequenzbereichsverfahren zur einkanaligen spektralen Sprachsignalentstörung könnten in vier Klassen aufgeteilt werden. Die älteste Klasse vertreten dabei Verfahren der spektralen Subtraktion, die auf der Additivität der Störung beruht und von daher sehr intuitiv, einfach und weit verbreitet ist. So wird in einem der früheren Verfahren in [Sch65] für die Entstörung von Kommunikationssignalen vorgeschlagen, spektrale Amplituden des Störsignals zu ermitteln, um sie anschließend vom Spektrum des Eingangssignals zu subtrahieren. Als namensgebende Arbeit dieser Klasse, die sich explizit auf Verarbeitung digitaler Sprachsignale im Frequenzbereich bezieht, können allerdings die Publikationen [Bol79] und [BSM79] angesehen werden. Bei Entwicklung moderner Verfahren zur spektralen Subtraktion werden sowohl das Sprach- als auch das Störspektrum als Zufallsprozesse modelliert und die SCHÄTZUNG DER SPEKTRALEN RAUSCHLEISTUNGSDICHTE wird zum unverzichtbaren Baustein solcher Systeme [TTM+11]. Mit Hilfe dieser Schätzung wird das Spektrum des gestörten Signals subtraktiv entstört, so dass die Operation der spektralen Subtraktion als Entstörungsregel eindeutig im Mittelpunkt der Algorithmen dieser Klasse steht.

Die zweite Klasse der Algorithmen der spektralen Entstörung wird als modellbasierte statistische Spektralentstörung bezeichnet, denn die Entstörungsregel beruht hier auf der statistischen Inferenz bzgl. des Spektrums des ungestörten Sprachsignals [Loi13]. In der Regel wird die Störung hier nicht subtraktiv sondern multiplikativ entfernt, wofür die BERECHNUNG DER SPEKTRALEN FILTERFUNKTION (engl. *gain function*) notwendig ist. Die Multiplikation der Filterfunktion mit dem Spektrum des gestörten Signals ergibt das Spektrum des entstörten Signals, dessen Rücktransformation in den Zeitbereich das entstörte Sprachsignal ergibt. Im Unterschied zur spektralen Subtraktion ist der bestimmende Parameter der Filterfunktion hier nicht die spektrale Rauschleistungsdichte (wenn überhaupt), sondern das sogenannte *a priori* Signal-zu-Rausch Verhältnisses (engl. *signal-to-noise ratio*, SNR) [2] [Cap94]. Die A PRIORI SNR-SCHÄTZUNG ist von daher ein fester Bestandteil der Algorithmen der modellbasierten Spektralentstörung. Begründet in der Dünnbesetztheit (engl. *sparseness*) der Sprachsignalspektren wird hier auch oft die Sprachanwesenheitswahrscheinlichkeit in die Berechnung der Filterfunktion integriert [Coh01]. An dieser Stelle werden Verfahren zur BERECHNUNG DER SPRACHPRÄSENZWAHRSCHEINLICHKEIT eingesetzt.

Die dritte Klasse der Algorithmen zur einkanaligen spektralen Sprachsignalentstörung stellen Algorithmen mit binären Masken dar. Während Verfahren der beiden ersten Klassen eine gute auditive Qualität der entstörten Signale bezwecken, zielen diese Algorithmen auf Verbesserung ihrer Verständlichkeit ab [Loi13]. Wie der Name bereits sagt, werden für die Spektralentstörung statt wertekontinuierlicher Filterfunktionen binäre Masken verwendet, die nur aus Nullen und Einsen bestehen. Im Unterschied zu Sprachpräsenzwahrscheinlichkeit, die aus einer weichen Entscheidung (engl. *soft decision*) entsteht, beruhen binären Masken auf einer binären Entscheidung (engl. *hard decision*).

In der vierten Klasse von Algorithmen werden tiefe neuronale Netze (engl. *deep neural networks*, DNN) eingesetzt [WN99, HCS+07]. Diese benötigen zwar eine aufwendige Trainingsphase mit vielen Stunden von Audiomaterial, um die Millionen ihrer Parameter zu trainieren, weisen dafür aber die beste Leistungsfähigkeit in der Signalentstörung im Vergleich zu den Verfahren der drei zuvor erwähnten Klassen von Algorithmen auf [XDDL14]. Die DNN-basierten Algorithmen sind eher datengetrieben als modellbasiert. Dabei werden DNNs nicht nur direkt zur Sprachsignalentstörung verwendet mit dem gestörten Sprachsi-

[2] Das *a priori* SNR wird im Zeit-Frequenz Bereich als Quotient der spektralen Leistungsdichte des ungestörten Sprachsignals zu der des Störsignals definiert.

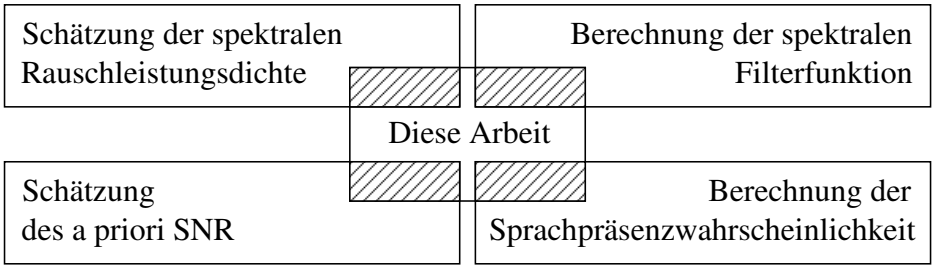


Abbildung 1.1.: Beitrag der Arbeit zur einkanaligen Sprachsignalentstörung.

gnal als Eingang und dem entstörten als Ausgang, sondern können auch Aufgaben verschiedener Verfahren zur einkanaligen spektralen Sprachsignalentstörung aus Abb. 1.1 übernehmen oder unterstützen sie in bestimmten Teilaufgaben [MT16]. Dabei entstehen Hybridsysteme, die sowohl modellbasierte als auch datengetriebene Verfahren kombinieren. Die DNN-basierte Entstörung von Sprachsignalen ist momentan ein attraktiver Gegenstand aktueller wissenschaftlicher Forschung.

Wie in Abb. 1.1 dargestellt werden im Rahmen dieser Arbeit modellbasierte Verfahren entwickelt, die in allen vier vorgestellten Klassen der Algorithmen zur einkanaligen spektralen Sprachsignalentstörung eingesetzt werden können. Diese Verfahren bilden vier grundlegende Bausteine eines Systems zur spektralen Sprachsignalentstörung:

- SCHÄTZUNG DER SPEKTRALEN RAUSCHLEISTUNGSDICHTE,
- BERECHNUNG DER SPEKTRALEN FILTERFUNKTION,
- SCHÄTZUNG DES A PRIORI SNR,
- BERECHNUNG DER SPRACHPRÄSENZWAHRSCHEINLICHKEIT.

Diese Arbeit besteht aus drei großen Teilen. In Teil A werden in Kap. 2 die Grundlagen der spektralen Sprachsignalentstörung vorgestellt, in Kap. 3 die konventionellen Verfahren beschrieben und in Kap. 4 die wissenschaftlichen Ziele dieser Arbeit definiert. Der Teil B ist der SCHÄTZUNG DER SPEKTRALEN RAUSCHLEISTUNGSDICHTE gewidmet. Zunächst werden zwei eigenständige Rauschschätzer vorgestellt. Der erste aus Kap. 5 stellt sich als ein *Minimum Statistics* Verfahren mit einer alternativen Steuerungsfunktion dar. Der zweite aus Kap. 6 basiert auf einem neuronalen Netz, das eine robuste spektrale Maske für die Abwesenheit eines ungestörten Sprachsignals berechnet. Anschließend wird in Kap. 7 ein Bayesscher Postprozessor präsentiert, der einem beliebigen Schätzer nachgeschaltet werden kann, um seine Schätzung zu präzisieren und so zu entstörten Signalen mit besserer Qualität zu gelangen. In Teil C werden drei Verfahren zur generalisierten spektralen Entstörung von Sprachsignalen präsentiert. Zunächst wird in Kap. 8 DIE BERECHNUNG DER SPEKTRALEN FILTERFUNKTION für die Entstörung von logarithmischen generalisierten spektralen Amplituden vorgestellt, die sich im Vergleich zu den modernen Filterfunktionen als konkurrenzfähig erweist. Anschließend wird in Kap. 9 das *Decision-Directed* Verfahren generalisiert, das für die SCHÄTZUNG DES A PRIORI SNR sehr häufig eingesetzt wird. Als nächstes wird in Kap. 10 ein Verfahren zur BERECHNUNG DER SPRACHPRÄSENZWAHRSCHEINLICHKEIT präsentiert, das Korrelationen zwischen den benachbarten Zeit-Frequenz-Punkten berücksichtigt, die in Sprachsignalen vorhanden sind. Und ganz zum Schluss wird in Kap. 11 eine ausführliche Zusammenfassung aller in dieser Arbeit entwickelten Verfahren gegeben.

A. GRUNDLAGEN, FORSCHUNGSSTAND UND ARBEITSZIELE

2. Spektrale Entstörung von einkanaligen Sprachsignalen

In diesem Kapitel werden Grundlagen der modellbasierten spektralen Entstörung von einkanaligen Sprachsignalaufnahmen erläutert. Dabei wird zunächst auf die statistische Modellierung der Sprachsignale und auf die dabei getroffenen Annahmen eingegangen. Anschließend wird die Transformation der Sprachsignale aus dem Zeitbereich in den Frequenzbereich und zurück mittels STFT eingeführt. Darauf basierend wird ein Analyse-Modifikation-Synthese (AMS) Framework betrachtet mit seinen vier grundlegenden Bausteinen: Schätzung des Rauschleistungsdichtespektrums (RLDS), Berechnung spektraler Filterfunktion, Schätzung des *a priori* SNR und Berechnung der Sprachpräsenzwahrscheinlichkeit (engl. *speech presence probability*, SPP). Im Weiteren werden Bewertungsmaße kurz beschrieben, die für die Bewertung der spektralen Sprachsignalentstörung benutzt werden. Dieses Grundlagenkapitel schließt mit der Beschreibung der Datenbanken, die im Rahmen dieser Arbeit für die Untersuchung der Systeme zur spektralen Entstörung von einkanaligen Sprachsignalen verwendet werden.

2.1. Analyse-Modifikation-Synthese System

Sei $y(n)$ ein einkanaliges zeitdiskretes gestörtes Sprachsignal, das sich als additive Überlagerung eines ungestörten Sprachsignals $s(n)$ und eines Störsignals $d(n)$ darstellt:

$$y(n) = s(n) + d(n). \tag{2.1}$$

Die Bezeichnung des ungestörten Sprachsignals als $s(n)$ und des Störsignals als $d(n)$ ist auf die englischen Begriffe *speech* und *distortion* zurückzuführen. Dabei ist $n \in \mathbb{Z}$ ein ganzzahliger Zeitindex. Die Signale $s(n)$ und $d(n)$ werden als Realisierungen zeitdiskreter nichtstationärer reellwertiger mittelwertfreier Zufallsprozesse modelliert, die voneinander statistisch unabhängig sind [CBHD08, Loi13].

In vielen Anwendungen der digitalen Sprachsignalverarbeitung ist das ungestörte Sprachsignal $s(n)$ von Interesse, obwohl nur das gestörte Sprachsignal $y(n)$ beobachtet wird. In diesen Anwendungen können Systeme zur Sprachsignalentstörung zum Einsatz kommen, welche eine Schätzung des ungestörten Sprachsignals $\hat{s}(n)$ aus $y(n)$ berechnen. Die Verwendung des Begriffes der Sprachsignalentstörung deutet darauf hin, dass diese Systeme prinzipiell bezwecken das Störsignal $d(n)$ aus dem gestörten Sprachsignal $y(n)$ zu entfernen, ohne dabei im prozessierten Signal $\hat{s}(n)$ künstliche Artefakte zu erzeugen oder das Sprachsignal zu verzerren. Der Weg zum Ziel führt häufig zum Kompromiss zwischen guter Störunterdrückung, Vermeiden von Artefakten und geringer Verzerrung des Sprachsignals [MM80]. Die Natürlichkeit des Restrauschens soll auch beachtet werden [SA05].

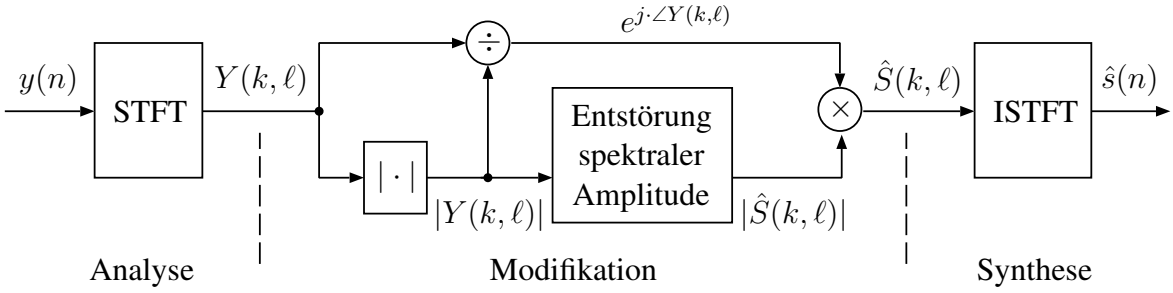


Abbildung 2.1.: Ein System zur einkanaligen spektralen Sprachsignalentstörung.

Wie in Kap. 1 erwähnt wird im Rahmen dieser Arbeit die Sprachsignalentstörung im Zeit-Frequenz-Bereich durchgeführt, der für Verarbeitung von nichtstationären Zufallsprozessen wie Sprachsignale gut geeignet ist [Mar03]. Dafür wird das gestörte Sprachsignal $y(n)$ in einem Analyse-Modifikation-Synthese Framework verarbeitet [CR83], indem es zunächst mit Hilfe der STFT in den Zeit-Frequenz-Bereich transformiert (Analyse), dort mit den Verfahren einer einkanaligen spektralen Entstörung verarbeitet (Modifikation) und anschließend mit Hilfe der inversen STFT (ISTFT) zurück in den Zeitbereich überführt wird (Synthese), siehe Abb. 2.1.

Analyse. Um die STFT-Koeffizienten von $y(n)$ zu berechnen, wird das Signal in überlappende Blöcke (oder Rahmen) der Länge $K \in \mathbb{N}_{>1}$ zerlegt. Um den Leakage-Effekt zu mildern, werden Abtastwerte eines jeden Signalblocks mit einem genauso langen Analyse-Fenster $\omega_{\mathrm{a}}(n)$ multipliziert, bevor auf das Produkt die diskrete Fourier-Transformation (DFT) angewandt wird, so dass für die STFT-Koeffizienten (oder Kurzzeitspektren) resultiert

$$y(n) \quad \overset{STFT}{\circ\!\!-\!\!\bullet} \quad Y(k,\ell) = \sum_{n=\ell\cdot R}^{\ell\cdot R+K-1} y(n)\cdot\omega_{\mathrm{a}}(n-\ell\cdot R)\cdot e^{-j\cdot\frac{2\pi k}{K}\cdot(n-\ell\cdot R)}, \tag{2.2}$$

wobei $k \in \{0,1,...,K-1\}$ der ganzzahlige Frequenzindex, K die DFT-Länge, $\ell \in \mathbb{N}$ der Blockindex, $R \in \mathbb{N}_{>0}$ der Rahmenvorschub (oder Dezimationsfaktor) und j die imaginäre Einheit sind. Man beachte, dass in dieser Arbeit die Länge der überlappenden Signalblöcke, die Länge des Fensters $\omega_{\mathrm{a}}(n)$ und die DFT-Länge gleich groß gewählt werden.

Um Vorteile der schnellen FFT bei der spektralen Signalentstörung nutzen zu können, wählt man häufig die Fensterlänge K zu einer Zweierpotenz. Obwohl dies eigentlich nicht zwingend erforderlich ist, wird oft auch der Rahmenvorschub R zur Zweierpotenz gewählt. Außerdem, damit das transformierte Signal aus seinen STFT-Koeffizienten perfekt rekonstruiert werden kann, muss für den Rahmenvorschub die Bedingung $R < K$ erfüllt werden [CR83]. In diesem Fall ist der Verschiebungsfaktor $a_R = \frac{R}{K} \in (0,1)$, der als ein Parameter der STFT verwendet wird, eine Potenz von $\frac{1}{2}$. Alternativ wird in der Fachliteratur ein Überlappungsfaktor in Prozent $\frac{1-R}{K}\cdot 100\,\%$ vorgegeben, der eine relative Anzahl der gemeinsamen Abtastwerte der aufeinander folgenden Signalblöcke angibt. Die typischen Werte für ein Überlappungsfaktor sind $50\,\%$ oder $75\,\%$, die jeweils einem Verschiebungsfaktor a_R von $\frac{1}{2}$ und $\frac{1}{4}$ entsprechen [GBM08].

Bei der Wahl der FFT-Länge K spielen nicht nur statistische spektrale Eigenschaften eines Sprachsignals eine Rolle sondern auch die Abtastrate. Während die untere Grenze für K durch die ausreichende Auflösung der Frequenzachse bestimmt wird, die für eine gute Signalentstörung benötigt wird, ist die obere Grenze durch die Zeitdauer begrenzt, in der ein

Sprachsignal noch als stationär (oder quasi-stationär) angenommen werden kann. Diese Zeitdauer wird in der Fachliteratur mit mit 15-30 ms in [RBQ+08] bzw. 20-40 ms in [EM84] angegeben. Um sie in die Blocklänge K umrechnen zu können, muss die Abtastrate des Sprachsignals bekannt sein. Die Untersuchungen bzgl. der Wahrnehmung von Sprachsignalen in Telefonie zeigten, dass ein Frequenzbereich zwischen 100 Hz und 9 kHz von Belang ist [FG50]. Ähnlich wird auch in [SA05] erwähnt, dass in einem Sprachsignal kaum Sprachsignalanteile unterhalb der Frequenz von 100 Hz vorhanden sind. Demnach sollen die mit 16 kHz abgetasteten Sprachsignale fast alle für die Wahrnehmung wichtige Informationen erfasst haben. Für die Abtastrate von 16 kHz korrespondieren die Fensterlängen von $K = \{2^8, 2^9\}$ zum Einen mit den Zeitdauern eines Signalblocks von jeweils $\{16, 32\}$ms, so dass ein Sprachsignal innerhalb solcher Signalausschnitte als stationär angesehen werden kann, und zum Anderen mit den Auflösungen der Frequenzachse von jeweils $\{62.5, 31.25\}$Hz. Mit diesen STFT-Parametern kann die Störung sogar zwischen den Harmonischen der Vokale unterdrückt werden, denn die mittlere Grundfrequenz der Sprachsignale der männlichen Sprecher beträgt in etwa 120 Hz und der weiblichen 200 Hz [TJB72, BCR75].

Modifikation. Die STFT-Koeffizienten $Y(k, \ell)$ besitzen einige Eigenschaften, die für spektrale Signalentstörung relevant sind. Da das gestörte Sprachsignal $y(n)$ und das Analysefenster $\omega_{\text{a}}(n)$ reellwertig sind, weisen die $Y(k, \ell)$ eines jeden ℓ-ten Blocks eine hermitesche Symmetrie um den Frequenzindex der Nyquist-Frequenz $k_{\text{Nyq}} = \frac{K}{2}$ auf (K als Zweierpotenz und somit als gerade Zahl angenommen). Von daher müssen nur die STFT-Koeffizienten der Frequenzbänder $k \in \{0,\, 1\,, \cdots \frac{K}{2}\}$ entstört werden, bevor sie für die Transformation zurück in den Zeitbereich entsprechend zusammengesetzt werden. Außerdem sind die STFT-Koeffizienten für $k = 0$ (der Gleichanteil) und für $k = k_{\text{Nyq}}$ reellwertig. Da die STFT (2.2) linear ist, gilt die Additivität (2.1) auch im Zeit-Frequenz-Bereich

$$Y(k, \ell) = S(k, \ell) + D(k, \ell), \tag{2.3}$$

wobei $S(k, \ell)$ und $D(k, \ell)$ jeweils die STFT-Koeffizienten des ungestörten Sprachsignals $s(n)$ und des Störsignals $d(n)$ darstellen. Die Additivität (2.3) dient als wichtiger Ausgangspunkt für Entwicklung von Entstörungsalgorithmen der spektralen Subtraktion, bei denen im Rahmen des Modifikation-Schrittes die Störung von der aktuellen Beobachtung $Y(k, \ell)$ subtraktiv entfernt wird. Außerdem erweist sich (2.3) als sehr nützlich sowohl für stochastische Modellierung von $Y(k, \ell)$ als auch für statistische Inferenz, die für spektrale Entstörung in den modellbasierten Algorithmen oft eingesetzt wird.

Eine wichtige Annahme, die bei stochastischer Modellierung eines STFT-Koeffizienten $S(k, \ell)$ getroffen wird, ist die statistische Unabhängigkeit zwischen seinem Realteil und seinem Imaginärteil. Diese Annahme folgt daraus, dass Betrag und Phase von $S(k, \ell)$ als statistisch unabhängig angenommen werden [EHHJ07]. Da $S(k, \ell)$ laut (2.2) als eine lineare Kombination von K aufeinander folgenden Abtastwerten eines zeitdiskreten Sprachsignals $s(n)$ berechnet wird, wird der zentrale Grenzwertsatz für die statistische Modellierung oft herangezogen [EM84]. Demnach unterliegen sowohl Real- als auch Imaginärteil von $S(k, \ell)$ asymptotisch eigentlich nur dann einer reellwertigen mittelwertfreien Normalverteilung, wenn alle Summanden in (2.2) statistisch unabhängig sind und K groß genug ist [PP02]. Allerdings wird die erste Bedingung der statistischen Unabhängigkeit bei den Sprachsignalen aufgrund vorhandener Korrelationen zwischen den aufeinander folgenden Abtastwerten verletzt. Dennoch darf der zentrale Grenzwertsatz laut [Bri74] auch bei den Signalen angewandt werden, deren aufeinander folgenden Abtastwerte statistisch schwach

abhängig (engl. *weakly dependent*) sind. Demnach kann man für Sprachsignale annehmen, dass $S(k,\ell)$ einer komplexwertigen mittelwertfreien Normalverteilung folgen [Loi13]. Experimentelle Untersuchungen wie in [JBHH05] untermauern diese Annahme. Auf ähnliche Weise wird auch statistische Modellierung von STFT-Koeffizienten eines Störsignals $D(k,\ell)$ begründet und zwar unabhängig von der Art der Hintergrundstörung. Somit können Verteilungsdichtefunktionen von $S(k,\ell)$ und $D(k,\ell)$ mit Hilfe von komplexwertigen mittelwertfreien Normalverteilungen modelliert werden

$$p_{S(k,\ell)}(s) = \mathcal{N}_{\mathbb{C}}(s;\, \lambda_S(k,\ell)), \qquad (2.4) \qquad p_{D(k,\ell)}(d) = \mathcal{N}_{\mathbb{C}}(d;\, \lambda_D(k,\ell)) \qquad (2.5)$$

mit den frequenzabhängigen zeitvarianten spektralen Leistungsdichten als Parameter

$$\lambda_S(k,\ell) \triangleq \mathbb{E}\left[|S(k,\ell)|^2\right], \qquad (2.6) \qquad \lambda_D(k,\ell) \triangleq \mathbb{E}\left[|D(k,\ell)|^2\right], \qquad (2.7)$$

die der Instationarität der Zufallsprozesse die Rechnung tragen. Unterdessen werden mit $\mathcal{N}_{\mathbb{C}}(\cdot)$ und $\mathbb{E}[\cdot]$ jeweils eine komplexwertige Normalverteilung und ein Erwartungswertoperator bezeichnet. Die Normalverteilung einer mittelwertfreien komplexwertigen Zufallsvariable $X \in \mathbb{C}$ mit der Varianz $\lambda_X \in \mathbb{R}_{>0}$ an der Stelle $X = x$ ist dabei wie folgt definiert:

$$\mathcal{N}_{\mathbb{C}}(x;\lambda_X) \triangleq \frac{1}{\pi\lambda_X} \cdot \exp\left(-\frac{|x|^2}{\lambda_X}\right). \qquad (2.8)$$

Die vorgestellte statistische Modellierung von $S(k,\ell)$ und $D(k,\ell)$ eingesetzt in (2.3) führt dazu, dass $Y(k,\ell)$ auch einer mittelwertfreien komplexwertigen Normalverteilung folgt:

$$p_{Y(k,\ell)}(y) = \mathcal{N}_{\mathbb{C}}(y;\lambda_Y(k,\ell)) \quad \text{mit} \quad \lambda_Y(k,\ell) = \lambda_S(k,\ell) + \lambda_D(k,\ell). \qquad (2.9)$$

Somit kann eine statistische Inferenz der STFT-Koeffizienten des ungestörten Sprachsignals $E[S(k,\ell)|Y(k,\ell)]$ berechnet werden, die zum linearen spektralen Wiener-Filter führt [LO79, VM06, Ast10], das in der modellbasierten spektralen Sprachsignalentstörung immer noch häufig eingesetzt wird. Das Wiener-Filter entfernt die Störung multiplikativ mit Hilfe einer reellwertigen spektralen Filterfunktion. Die Tatsache, dass die spektrale Filterfunktion reellwertig ist, führt dazu, dass die Phase von $Y(k,\ell)$ als die Phase der STFT-Koeffizienten des entstörten Sprachsignals übernommen wird. Mit anderen Worten ist die gestörte Phase der optimale Schätzer für die Phase des entstörten Signals im *minimum mean square error* (MMSE) Sinn [EM84].

Dieses Ergebnis geht mit den Untersuchungen einher, dass der spektralen Amplitude häufig eine höhere Bedeutung für die wahrnehmbare Signalqualität beigemessen wird [WL82], obwohl die Phase der STFT-Koeffizienten eines Sprachsignals auch eine bedeutende Rolle für die Signalsynthese spielen kann [PWS11]. Aus diesem Grund konzentriert sich diese Arbeit auf der Entstörung spektraler Amplituden und lässt die Phase des gestörten Signals das System unverändert passieren, so dass das Kurzzeitspektrum $\hat{S}(k,\ell)$ des entstörten Signals wie folgt zusammengesetzt wird:

$$\hat{S}(k,\ell) = |\hat{S}(k,\ell)| \cdot e^{j\cdot\angle Y(k,\ell)} \quad \text{mit} \quad e^{j\cdot\angle Y(k,\ell)} = \frac{Y(k,\ell)}{|Y(k,\ell)|}. \qquad (2.10)$$

Wie man in Abb. 2.1 sieht, wird im AMS-System zur spektralen Sprachsignalentstörung die Phase des gestörten Sprachsignals als Phase des entstörten Signals verwendet. Dies ist allerdings nicht immer der Fall und es gibt Verfahren wie [GKR12] oder [MSS16], welche die

Phase des entstörten Sprachsignals explizit schätzen. Im Rahmen dieser Arbeit wird darauf allerdings verzichtet.

Synthese: Die Rekonstruktion des entstörten zeitdiskreten Sprachsignals $\hat{s}(n)$ aus seinen STFT-Koeffizienten $\hat{S}(k,\ell)$ wird mit Hilfe der gewichteten *overlapp-add* (OLA) Methode realisiert [Cro80], die auch als gleitende diskrete Fouriertransformation bekannt ist:

$$\hat{s}(n) = \sum_{\ell=0}^{L-1} \sum_{k=0}^{K-1} \hat{S}(k,\ell) \cdot \omega_{\mathrm{s}}(n - \ell \cdot R) \cdot e^{j \cdot \frac{2\pi k}{K} \cdot (n - \ell \cdot R)}. \tag{2.11}$$

$\omega_{\mathrm{s}}(n)$ ist dabei ein reellwertiges Synthese-Fenster. Man beachte, dass das Analyse- und das Synthese-Fenster die Bedingung der perfekten Rekonstruktion (engl. *completeness condition*) für jeden ℓ-ten Signalblock erfüllen müssen [AC07]

$$\sum_{\ell'=\ell-\frac{K}{R}}^{\ell+\frac{K}{R}} w_a(n - \ell' \cdot R) \cdot w_s(n - \ell' \cdot R) = 1 \quad \text{für} \quad 0 \leq n \leq K-1, \tag{2.12}$$

wobei das Verhältnis $\frac{K}{R} \in \mathbb{N}_{>1}$ als eine ganze Zahl angenommen wurde. Die Fensterfunktionen, welche die Bedingung (2.12) erfüllen, sind auch biorthogonal zueinander [CB01b]. Da bei einem gegebenen Analyse-Fenster die Wahl des Synthese-Fensters, das die Bedingung (2.12) erfüllt, nicht eindeutig ist, wird das Synthese-Fenster verwendet, das die euklidische Distanz zwischen den beiden Fenstern $\omega_{\mathrm{a}}(n)$ und $\omega_{\mathrm{s}}(n)$ minimiert [WR90].

Nachdem das gesamte AMS-System zur einkanaligen spektralen Sprachsignalentstörung aus Abb. 2.1 erläutert wurde, soll im Weiteren auf das Herzstück des Systems – die Entstörung spektraler Amplituden näher eingegangen werden.

2.2. Bausteine der Entstörung spektraler Amplituden

In Abb. 2.2 ist ein Blockschaltbild eines konventionellen Systems zur Entstörung spektraler Amplituden mit seinen vier grundlegenden Bausteinen dargestellt:

- Baustein 1: Schätzung des Rauschleistungsdichtespektrums
- Baustein 2: Berechnung der spektralen Filterfunktion
- Baustein 3: A priori SNR-Schätzung
- Baustein 4: Berechnung der Sprachpräsenzwahrscheinlichkeit.

Zwei dieser Bausteine nämlich die RLDS-Schätzung und die Berechnung einer spektralen Filterfunktion sind unabdingbar für die spektrale Entstörung[1]. So kann mit diesen z. B. eine spektrale Leistungssubtraktion (SL) realisiert werden [CN78, SF96]. Möchte man eine spektrale Filterfunktion für die Entstörung verwenden, die von dem *a priori* SNR abhängt (z. B. das Wiener-Filter) [LO79, Ast10], muss zusätzlich die *a priori* SNR-Schätzung im System realisiert sein. Außerdem gibt es Systeme, in denen die spektrale Filterfunktion von der

[1] In dem Fall, wenn ein neuronales Netz die Aufgabe der Signalentstörung allein übernimmt wie in [XDDL14], gilt diese Aussage natürlich nicht.

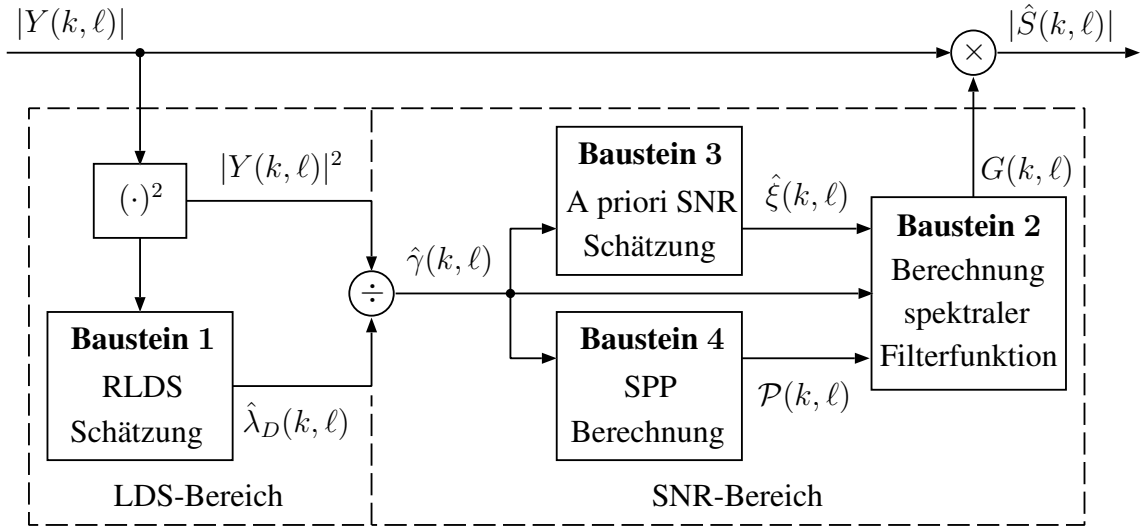


Abbildung 2.2.: Grundbausteine eines Systems zur Entstörung spektraler Amplituden.

Sprachpräsenzwahrscheinlichkeit abhängt [MM80,Coh01], die in diesem Fall auch geschätzt werden muss. Während der Baustein 1 im Bereich des Leistungsdichtespektrums (LDS) realisiert wird, arbeiten die Bausteine 2-4 häufig im Bereich der relativen Größen wie das SNR-Bereich[2], wie die weiteren Ausführungen zeigen werden.

Baustein 1 - Schätzung des Rauschleistungsdichtespektrums: Als eine der ersten Methoden zur Sprachsignalentstörung im Frequenzbereich wurde die spektrale Leistungssubtraktion entwickelt, die zunächst in [Sch65] und dann in [CN78] beschrieben wird. Dabei geht man von der Annahme der Additivität der momentanen spektralen Leistungen aus

$$|Y(k,\ell)|^2 = |S(k,\ell)|^2 + |D(k,\ell)|^2, \tag{2.13}$$

die im Unterschied zur Gleichung $E[|Y(k,\ell)|^2] = E[|S(k,\ell)|^2] + E[|D(k,\ell)|^2]$, die auf Unkorreliertheit von $S(k,\ell)$ und $D(k,\ell)$ basiert, nur eine Näherung ist. Laut [LL08] ist sie allerdings für kleine und große Werte des globalen SNR im gestörten Signal durchaus gutmütig. Da in Realität die momentane Rauschleistung $|D(k,\ell)|^2$ nicht gegeben ist, wird sie bei spektraler Leistungssubtraktion häufig durch einen RLDS-Schätzwert $\hat{\lambda}_D(k,\ell) \in (0;\, |Y(k,\ell)|^2)$ ersetzt. Somit ergibt sich eine einfache jedoch nichtlineare Entstörungsregel:

$$|\hat{S}_{\text{SL}}(k,\ell)| = \sqrt{|Y(k,\ell)|^2 - \hat{\lambda}_D(k,\ell)}. \tag{2.14}$$

Da das RLDS $\lambda_D(k,\ell)$ aus (2.7) im Bereich der spektralen Leistungen definiert ist, wird die spektrale Amplitude des gestörten Signals quadriert, bevor sie an den Eingang des RLDS-Schätzers weitergeleitet wird. Die Aufgabe des RLDS-Schätzers besteht nun darin, $\lambda_D(k,\ell)$ allein aus der aktuellen (und den vergangenen) Beobachtung $|Y(k,\ell)|^2$ zu schätzen, die oft als eine beobachtbare exponentialverteilte Zufallsvariable aufgefasst wird. Und da $\lambda_D(k,\ell)$ sich als Parameter der Verteilungsdichtefunktionen der Zufallsvariable $|Y(k,\ell)|^2$ darstellt, kann die Rauschschätzung als Parameterschätzung angesehen werden. Außerdem werden bei der Schätzung von $\lambda_D(k,\ell)$ solche Eigenschaften der akustischen Signale ausgenutzt, wie die

[2]Es gibt Verfahren, welche die Bausteine 2-4 auch in anderen Bereichen als dem SNR-Bereich realisieren. Allerdings lassen sie sich häufig für den Einsatz im SNR-Bereich umformulieren wie z. B. [EMTF15].

Dünnbesetztheit der Kurzzeitspektren des Sprachsignals, die Omnipräsenz des Rauschens oder die Tatsache, dass die STFT-Koeffizienten benachbarter Zeit-Frequenz-Punkte korreliert sind. Näher wird darauf in Abschnitt 3.1 eingegangen. An diese Stelle sei nur erwähnt, dass eine RLDS-Schätzung in Gegenwart einer nichtstationären Störung besonders herausfordernd ist [RL06]. Während eine Überschätzung der spektralen Rauschleistungsdichte zur unerwünschten Verzerrung des Sprachsignals und zum daraus resultierenden Verlust von Qualität und Verständlichkeit des entstörten Signals führt, resultiert eine Unterschätzung des RLDS in einem unangenehmen restlichen Rauschen im entstörten Signal, dem sogenannten musikalischen Rauschen (engl. *musical noise*) [HJH08].

Alternativ kann die Leistungssubtraktion (2.14) auch wie folgt umgeschrieben werden:

$$|\hat{S}_{\text{SL}}(k,\ell)| = \sqrt{\frac{\hat{\gamma}(k,\ell)-1}{\hat{\gamma}(k,\ell)}} \cdot |Y(k,\ell)|, \tag{2.15}$$

wobei $\hat{\gamma}(k,\ell)$ ein Schätzwert des sogenannten *a posteriori* SNR ist[3], das als Verhältnis der momentanen spektralen Leistung $|Y(k,\ell)|^2$ zur RLDS $\lambda_D(k,\ell)$ definiert ist

$$\gamma(k,\ell) \triangleq \frac{|Y(k,\ell)|^2}{\lambda_D(k,\ell)}. \tag{2.16}$$

Während im Allgemeinen das *a posteriori* SNR reellwertig und positiv $\gamma(k,\ell) \in \mathbb{R}_{>0}$ ist, darf sein Schätzwert bei der Verwendung in der spektralen Leistungssubtraktion (2.15) nicht kleiner 1 sein. Sonst fällt auf, dass bei der spektralen Leistungssubtraktion (2.15) weder eine *a priori* SNR-Schätzung noch eine Schätzung der *a posteriori* SPP benötigt werden.

Die Tatsache, dass das *a posteriori* SNR $\gamma(k,\ell)$ nicht nur bei spektraler Leistungssubtraktion verwendet wird, sondern auch in vielen anderen Filterfunktionen, führte dazu, dass $\gamma(k,\ell)$ sich in der spektralen Sprachsignalentstörung als eine eigenständige Größe etabliert hat. Da $\gamma(k,\ell)$ in einem System zur Entstörung spektraler Amplitude direkt nach der RLDS-Schätzung berechnet wird, kann ein Rauschschätzer auch als Schätzer des *a posteriori* SNR betrachtet werden.

Baustein 2 - Berechnung der spektralen Filterfunktion: Wie bei der Umformulierung der spektralen Leistungssubtraktion (2.15) angedeutet, ist es in der spektralen Signalentstörung üblich, eine Entstörungsregel mit Hilfe einer spektralen Filterfunktion anzugeben, die folgendermaßen definiert wird:

$$G(k,\ell) \triangleq \frac{|\hat{S}(k,\ell)|}{|Y(k,\ell)|}. \tag{2.17}$$

Also sind alle spektralen Filterfunktionen, die im Rahmen dieser Arbeit verwendet werden, reellwertig $G(k,\ell) \in \mathbb{R}_{>0}$. Allerdings kann $G(k,\ell)$ in einer konkreten Anwendung in einen Wertebereich $G(k,\ell) \in (G_{\text{min}},\, G_{\text{max}})$ eingegrenzt werden [EM84, Cap94, LL08].

[3] Die Bezeichnung von $\gamma(k,\ell)$ mit dem zunächst ungewöhnlichen lateinischen Begriff *a posteriori* kann wie folgt erklärt werden. Laut dem Rechtschreiblexikon von Duden gibt es zwei Bedeutungen von *a posteriori*. Die erste Bedeutung 'aus der Erfahrung gewonnen' weist darauf hin, dass $\gamma(k,\ell)$ nur berechnet werden kann, wenn man $|Y(k,\ell)|^2$ beobachtet hat (Wissen aus Erfahrung). Die zweite Bedeutung 'nachträglich oder später' signalisiert, dass $\gamma(k,\ell)$ zum Signal gehört, nachdem es vom Störsignal überlagert wurde. Die zweite Interpretation scheint dem Autor etwas sinnvoller als die erste zu sein.

Die Notwendigkeit der unteren Grenze G_{min} lässt sich am Beispiel der bereits eingeführten spektralen Leistungssubtraktion $G_{\text{SL}}(k,\ell)$ aus (2.15) gut erklären, die ja auf der Entstörungsregel (2.14) basiert. In den Zeit-Frequenz-Punkten mit einer sehr kleinen momentanen spektralen Leistung des Sprachsignals $|S(k,\ell)|^2 \ll |D(k,\ell)|^2$ weist die Amplitude der Kurzzeitspektren des entstörten Signals $|\hat{S}_{\text{SL}}(k,\ell)|$ verhältnismäßig hohe Varianz auf, die vermehrt zu den plötzlichen energiereichen Peaks führt, die in der Zeit-Frequenz-Ebene in etwa gleichmäßig verteilt sind. Dadurch entstehen im entstörten Signal unangenehme Prozessierungsartefakte – das sogenannte musikalische Rauschen (engl. *musical tones*) [VHH98]. Wie in [BSM79] vorgeschlagen, kann das musikalische Rauschen durch Zulassung einer kontrollierbaren Menge der spektralen Rauschleistung im entstörten Signal gemildert werden, denn die energiereichen Peaks versinken dabei im vorgegebenen Rauschpegel [Cap94][4]. Realisiert wird diese Idee durch Einführung der Untergrenze der Filterfunktion G_{min}.

Wird $G_{\text{ref}} = 1$ (keine spektrale Entstörung) als ein Referenzwert verwendet und die Filterfunktion als eine sogenannte Feldgröße[5] betrachtet wie in [EM84], kann G_{min} in Dezibel angegeben werden

$$G_{\text{min}}\,/\text{dB} \triangleq 20 \cdot \log_{10} \frac{G_{\text{min}}}{G_{\text{ref}}} = 20 \cdot \log_{10} G_{\text{min}}. \tag{2.18}$$

In der Fachliteratur sind unterschiedliche Angaben für G_{min} zu finden wie $-15\,\text{dB} \approx 0.18$ für eine modifizierte *maximum Likelihood* (ML) Filterfunktion in [Yan93], $-25\,\text{dB} \approx 0.06$ für eine MMSE-basierte *log-spectral amplitude* (LSA) Filterfunktion in [Coh01] oder sogar $-40\,\text{dB} = 0.01$ für eine MMSE-Filterfunktion der generalisierten spektralen Subtraktion in [LSH$^+$08]. Weitere Empfehlungen für G_{min} sind $(-20\,\text{dB};\ -10\,\text{dB}) \approx (0.1;\ 0.32)$ für das analytische Wiener-Filter in [Ars06] oder sogar $(-20\,\text{dB};\ -6\,\text{dB}) \approx (0.1;\ 0.5)$ für das ML-basierte Wiener-Filter in [RBKS14], das sich als Filterfunktion der spektralen Leistungssubtraktion $G^2_{\text{SL}}(k,\ell)$ herausstellt [MM80]. Bei der Wahl von G_{min} muss also unbedingt die Art der verwendeten Filterfunktion in Betracht gezogen werden. Grundsätzlich gilt folgende Regel: je kleiner die Menge an musikalischem Rauschen, die von einer spektralen Entstörung produziert wird, desto kleiner kann G_{min} gewählt werden. Kleinere Werte von G_{min} sind vorteilhaft für die Rauschunterdrückung und aus diesem Grund auch erstrebenswert.

Bei einer gegebenen Untergrenze G_{min} kann die resultierende Filterfunktion ähnlich wie in [GBM08] berechnet werden:

$$G(k,\ell) = \max\left(\, G_{\text{Art}}(k,\ell),\, G_{\text{min}} \,\right), \tag{2.19}$$

wobei $G_{\text{Art}}(k,\ell)$ die Filterfunktion bestimmter Art ist. Man beachte, dass $G_{\text{Art}}(k,\ell)$ häufig als $G_{H_1}(k,\ell)$ bezeichnet wird, wobei H_1 für eine Hypothese für eine Sprachsignalpräsenz steht. Ein entscheidender Punkt dabei ist, dass bei Herleitung spektraler Filterfunktionen generell von Sprachsignalpräsenz ausgegangen wird [MM80].

Bemerkenswert ist, dass die obere Grenze vieler spektraler Filterfunktion wie der spektralen Leistungssubtraktion oder des Wiener-Filters per definitionem den Wert $G_{\text{max}} = 1$ aufweist. Diese Tatsache spiegelt den Sachverhalt wieder, dass im Rahmen der spektralen

[4] Weitere Reduzierung des musikalischen Rauschens kann man durch eine zusätzliche Reinigung des Spektrogramms des entstörten Signals von scharfen energiereichen Gipfeln erreichen [GTT98]. Dafür werden zunächst die Zeit-Frequenz-Punkte mit musikalischem Rauschen gefunden und anschließend entfernt.

[5] Bevor eine Größe in dB angegeben wird, muss unbedingt kenntlich gemacht werden, ob sie entweder als eine Feld- oder eine Leistungsgröße definiert wird [Bla98].

Entstörung eigentlich eine Rauschunterdrückung stattfindet, so dass die spektralen Amplituden des verrauschten Signals $|Y(k,\ell)|$ gedämpft und nicht verstärkt werden, was für die Zeit-Frequenz-Punkte mit $|Y(k,\ell)| > |S(k,\ell)|$ durchaus sinnvoll ist. Allerdings in dem Fall, wenn spektrale Phase des Störsignals $\angle D(k,\ell)$ entgegengesetzt der Phase des ungestörten Signals $\angle S(k,\ell)$ ist, führt die additive Störung zur Ungleichung $|Y(k,\ell)| < |S(k,\ell)|$, so dass die Verstärkung der spektralen Amplitude $|Y(k,\ell)|$ eine bessere Wahl für Signalentstörung wäre, was auch die Werte der Filterfunktion $G(k,\ell) > 1$ rechtfertigt. So gibt es eine Reihe von Filterfunktionen, die sich analytisch auf den ganzen Wertebereich der positiven reellen Zahlen erstrecken [EM84, EM85, SF96, LL11]. Diese Filterfunktionen hängen meistens antiproportional von dem *a posteriori* SNR $\gamma(k,\ell)$ ab, so dass in einer Anwendung sehr große Werte der Filterfunktion vorkommen können, wenn der Schätzwert $\hat{\gamma}(k,\ell)$ zu klein geschätzt wird, d. h. wenn der Fall $|Y(k,\ell)|^2 \ll \hat{\lambda}_D(k,\ell)$ vorliegt. Aus diesem Grund macht eine Obergrenze für eine Filterfunktion Sinn, die im einfachsten Fall zum neutralen Element $G_{\text{max}} = 1$ gewählt werden kann [LL08, JH12].

Obwohl in der Sprachsignalentstörung die Werte $G_{\text{max}} > 1$ untypisch sind, werden in den Anwendungen zur Verbesserung der Verständlichkeit von Sprachsignalen die Filterfunktionen mit den vorgegebenen Werten G_{max} durchaus verwendet. Dabei werden die STFT-Koeffizienten der ungestörten Sprachsignale mit einer Filterfunktion so modifiziert, dass sie in einer verrauschten Umgebung bessere Sprachverständlichkeit aufweisen [SV06]. Dabei sind in der Fachliteratur die G_{max} Werte von $12\,\text{dB} \approx 4$ in [IS08], von $20\,\text{dB} = 10$ in [Bro09] und sogar von $30\,\text{dB} \approx 31.6$ in [SV06, ZKS12] zu finden.

Sind beide Grenzen der Filterfunktion G_{min} und G_{max} vorgegeben, wird die resultierende Filterfunktion ähnlich wie in [IS08] berechnet:

$$G(k,\ell) = \min\left(\max\left(G_{\text{min}},\, G_{\text{Art}}(k,\ell)\right),\, G_{\text{max}}\right). \tag{2.20}$$

In der Regel sind G_{min} und G_{max} Konstanten, die weder vom Frequenzindex k noch vom Blockindex ℓ abhängen. Der konstante Wert von G_{min} sorgt allerdings dafür, dass die Störung im entstörten Signal nur proportional gedämpft wird. Dies ist z. B. für die transienten kurzzeitigen Störungen nicht gut geeignet, denn in diesem Fall werden sie im entstörten Signal immer noch wahrnehmbar sein. Um dies zu verhindern, kann eine Unterschranke $G_{\text{min}}(k,\ell)$ verwendet werden, die sowohl vom Frequenzindex als auch vom Zeitindex abhängt [Pud99]. Außerdem kann $G_{\text{min}}(k,\ell)$ dafür benutzt werden, einen gewünschten Verlauf der Rauschleistungsdichte des im entstörten Signal zugelassenen Rauschens einzuprägen [RBKS14].

Baustein 3 - A priori SNR-Schätzung: Eine weitere Möglichkeit das musikalische Rauschen im prozessierten Sprachsignal zu bekämpfen, wird in [SF96] vorgeschlagen. Die Motivation dafür ist die große Varianz des *a posteriori* SNR-Schätzers $\hat{\gamma}(k,\ell)$ und die damit verbundene Unsicherheit, welche in die Berechnung von Filterfunktionen einfließt. Die Abhilfe soll die Ersetzung von $\gamma(k,\ell)$ durch seinen Erwartungswert $E[\gamma(k,\ell)]$ verschaffen, der unter Berücksichtigung von (2.16) berechnet werden kann:

$$\mathbb{E}[\gamma(k,\ell)] = \xi(k,\ell) + 1, \tag{2.21}$$

wobei $\xi(k,\ell) \in \mathbb{R}_{>0}$ das sogenannte *a priori* SNR ist, das als Verhältnis des LDS des ungestörten Sprachsignals $\lambda_S(k,\ell)$ aus (2.6) zum RLDS $\lambda_D(k,\ell)$ aus (2.7) ähnlich wie in [EM83] definiert wird

$$\xi(k,\ell) \triangleq \frac{\lambda_S(k,\ell)}{\lambda_D(k,\ell)}. \tag{2.22}$$

Man beachte, dass das *a priori* SNR $\xi(k,\ell)$ im Unterschied zum *a posteriori* SNR $\gamma(k,\ell)$ per definitionem ein Parameter und keine Zufallsvariable mehr ist. Ersetzt man in (2.15) das *a posteriori* SNR $\gamma(k,\ell)$ durch seinen Erwartungswert $\mathbb{E}[\gamma(k,\ell)]$ aus (2.21), resultiert eine Filterfunktion der modifizierten spektralen Leistungssubtraktion wie in [SF96]:

$$G_{\text{MSL}}(k,\ell) = \sqrt{\frac{\xi(k,\ell)}{\xi(k,\ell)+1}}. \tag{2.23}$$

Im Unterschied zur konventionellen Leistungssubtraktion aus (2.15) führt $G_{\text{MSL}}(k,\ell)$ zu einer linearen spektralen Filterung, denn $|Y(k,\ell)|$ fließt in die Berechnung von $G_{\text{MSL}}(k,\ell)$ nicht mehr ein. Laut den Untersuchungen in [SF96] führt die Verwendung von (2.23) statt (2.15) zu entstörten Signalen mit weniger Artefakten. Der Preis dafür ist die Notwendigkeit, das *a priori* SNR zu schätzen. Diese Aufgabe wird vom dritten Baustein eines Systems zur spektralen Entstörung übernommen. Bemerkenswert ist an dieser Stelle die Tatsache, dass die spektrale Filterfunktion (2.23) dem linearen Wiener-Filter sehr ähnelt [LO79, CG08], dessen spektrale Filterfunktion

$$G_{\text{WF}}(k,\ell) \triangleq \frac{\mathbb{E}[\,S(k,\ell)\,|\,Y(k,\ell)\,]}{|Y(k,\ell)|} = \frac{\xi(k,\ell)}{\xi(k,\ell)+1} \tag{2.24}$$

auch nur vom *a priori* SNR abhängt. Man beachte, dass (2.24) basierend auf der statistischen Modellierung hergeleitet werden kann, die in Abschnitt 2.1 vorgestellt wurde [VM06, Ast10]. Während die linearen Filterfunktionen im Bezug auf kleine Verzerrungen der gefilterten Sprachsignale vorteilhaft sind, weisen sie nur eine begrenzte Leistungsfähigkeit bzgl. Störunterdrückung auf. Außerdem behandeln sie nichtlineare Störungen und Störungen, die nicht Normalverteilt sind, ineffizient [HCS+07]. Aus diesen Gründen werden zunehmend Verfahren mit einer nicht-linearen Filterung eingesetzt. Diese Arbeit beschäftigt sich mit der Entwicklung von Verfahren zur einkanaligen spektralen Sprachsignalentstörung, die für eine nichtlineare Filterung eingesetzt werden.

Das wohl am weitesten verbreitete Verfahren zur *a priori* SNR-Schätzung ist das *Decision-Directed* (DD) Verfahren [EM84], das $\xi(k,\ell)$ aus dem *a posteriori* SNR $\hat{\gamma}(k,\ell)$ schätzt und in Abschnitt 3.3 näher beschrieben wird. Also baut das DD-Verfahren auf einer RLDS-Schätzung auf, die auch in dem Fall, wenn die Filterfunktion wie in (2.23) oder in (2.24) nur von $\xi(k,\ell)$ und nicht direkt von $\gamma(k,\ell)$ abhängt, unbedingt realisiert werden muss. Zu beachten ist auch, dass das DD-Verfahren laut [Cap94] auch unter einem 'leichten musikalischen Rauschen' leidet, dessen Milderung durch die Einführung einer unteren Grenze ξ_{min} möglich ist

$$\xi_{\text{out}}(k,\ell) = \max\left(\xi_{\text{in}}(k,\ell),\ \xi_{\text{min}}\right). \tag{2.25}$$

Dabei soll der Wert von ξ_{min} entsprechend dem mittleren *a priori* SNR der Zeit-Frequenz-Punkte gesetzt werden, in denen das Störsignal stark dominiert [Cap94]. Der Mechanismus der Milderung des musikalischen Rauschens durch die Einführung von ξ_{min} ist ähnlich wie bei der Einführung von G_{min} für die Filterfunktion. So können ξ_{min} und G_{min} bei Verwendung von Filterfunktionen (2.23) und (2.24) ineinander leicht umgerechnet werden. Dabei muss berücksichtigt werden, dass das *a priori* SNR eine Leistungsgröße ist, die durch eine Einführung des Referenzwertes $\xi_{\text{ref}} = 1$, der die Gleichung $\lambda_S(k,\ell) = \lambda_D(k,\ell)$ voraussetzt, in Dezibel angegeben werden kann:

$$\xi(k,\ell)\,/\,\text{dB} \triangleq 10\cdot\log_{10}\frac{\xi(k,\ell)}{\xi_{\text{ref}}} = 10\cdot\log_{10}\xi(k,\ell). \tag{2.26}$$

So wird in [Cap94] $\xi_{\text{min}} = -15\,\text{dB} \approx 0.03$ für den MMSE-Schätzer der spektralen Amplitude (SA) vorgeschlagen, die in [EM84] über den Erwartungswert $\mathbb{E}[\,|S(k,\ell)|\,|\,Y(k,\ell)\,]$ definiert ist. In [Coh04] wird $\xi_{\text{min}} = -25\,\text{dB} \approx 0.003$ für eine Reihe von Filterfunktionen gewählt. In [MWJ00] wird ξ_{min} für die Filterfunktion aus [MCA99] in Abhängigkeit vom globalen eingangsseitigen SNR mit einer Untergrenze von etwa $-9\,\text{dB}$ gewählt. Sonst sensibilisieren die Untersuchungen in [Yu13] dazu, den Parameter ξ_{min} in Abhängigkeit von der verwendeten Filterfunktion zu wählen. Der optimale Parameter ξ_{min} ist außerdem zweckgebunden und hängt davon ab, ob beispielsweise bei der Entstörung die Qualität der entstörten Signale oder die Störunterdrückung Priorität hat.

Die erwähnte MMSE-SA-Filterfunktion ist ein Beispiel für eine spektrale Filterfunktion, die nicht nur vom *a priori* SNR $\xi(k,\ell)$ sondern auch vom *a posteriori* SNR $\gamma(k,\ell)$ abhängig ist, unter denen $\xi(k,\ell)$ als dominanter Parameter und $\gamma(k,\ell)$ als ein Korrekturterm angesehen werden [Cap94]. Somit ist eine robuste Schätzung von $\xi(k,\ell)$ für eine erfolgreiche spektrale Entstörung von großer Bedeutung. Da weder $\lambda_S(k,\ell)$ noch $\lambda_D(k,\ell)$ in einer Anwendung bekannt sind, ist die Schätzung von $\xi(k,\ell)$ anspruchsvoller als die von $\gamma(k,\ell)$, zumal $\lambda_S(k,\ell)$ einen höheren Grad an Instationarität mit sich bringt als $\lambda_D(k,\ell)$ [FG14]. An dieser Stelle soll erwähnt werden, dass es durchaus üblich ist, die spektralen Filterfunktionen als Funktionen des *a priori* SNR und des *a posteriori* SNR anzugeben und nicht als Funktionen von $\lambda_D(k,\ell)$ und $\lambda_S(k,\ell)$. Diese Tatsache rechtfertigt die Einführung des SNR-Bereichs in Abb. 2.2 noch einmal.

Ein weiteres Beispiel für eine Filterfunktion, die sowohl von $\xi(k,\ell)$ als auch von $\gamma(k,\ell)$ abhängig sind, ist der MMSE-Schätzer der logarithmischen spektralen Amplitude (MMSE-LSA) aus [EM85] mit

$$G_{\text{LSA}}(k,\ell) \triangleq \frac{\exp\left(\,\mathbb{E}[\,\ln|S(k,\ell)|\,|\,Y(k,\ell)\,]\,\right)}{|Y(k,\ell)|} = G_{\text{WF}}(k,\ell)\cdot\exp\left(\frac{1}{2}\int_{v(k,\ell)}^{\infty}\frac{e^{-t}}{t}dt\right), \quad (2.27)$$

wobei die untere Grenze des Exponentialintegrals $v(k,\ell)$ entsprechend [Coh01] definiert ist:

$$v(k,\ell) \triangleq G_{\text{WF}}(k,\ell)\cdot\gamma(k,\ell) \quad (2.28)$$

Während der erste Term von (2.27) sich als ein Wiener-Filter darstellt und im Wertebereich $(0,\,1)$ liegt, kann die Integralexponentialfunktion durchaus Werte größer 1 in dem Fall annehmen, wenn $\gamma(k,\ell)$ und somit auch $v(k,\ell)$ zu klein wird. Also benötigt die MMSE-LSA-Filterfunktion sowohl die untere G_{min} als auch die obere Grenze G_{max}. Ähnliches gilt auch für die oben erwähnte MMSE-SA-Filterfunktion. Dabei sind beide Filterfunktionen MMSE-SA und MMSE-LSA durch ihre Abhängigkeit von $\gamma(k,\ell)$ nichtlinear, denn $|Y(k,\ell)|$ fließt in $\gamma(k,\ell)$ und somit auch in diese Filterfunktionen quadratisch ein. Die Untersuchungen in [HL06] zeigten, dass MMSE-LSA-Filterfunktion insgesamt zu den besten modellbasierten Filterfunktionen hinsichtlich der Qualität der prozessierten Signale gehört[6].

In einem System mit spektraler Entstörung mittels Filterfunktion (2.23), (2.24) und (2.27) müssen die Bausteine 1, 2 und 3 realisiert werden. Solche Systeme bekämpfen erfolgreich das musikalische Rauschen, vermeiden geschickt allzu große Verzerrungen des Sprachsignals und liefern somit die entstörten Signale mit guter Qualität. Der Preis für eine gute

[6] Neben modellbasierten Filterfunktionen, die aus einer statistischen Modellierung hergeleitet werden, können auch die sogenannten datengetriebenen Filterfunktionen experimentell berechnet werden, die zur etwas besseren Leistungsfähigkeit führen [EJH07].

Signalqualität ist allerdings eine mangelhafte Störunterdrückung, die konservativ und vorsichtig aufgrund der bereits erwähnten Tatsache ausfällt, dass spektrale Filterfunktionen unter der Annahme der Sprachsignalpräsenz hergeleitet werden [MM80].

Baustein 4 - Berechnung der Sprachpräsenzwahrscheinlichkeit: Wird die Sprachpräsenzwahrscheinlichkeit in die Berechnung der spektralen Filterfunktion integriert, lässt sich eine deutlich bessere Störsignaldämpfung erreichen [MCA99]. Die Untersuchungen wie in [SYMA05] haben gezeigt, dass $98\,\%$ der gesamten Energie eines Sprachsignals nur in etwa $9\,\%$ der Zeit-Frequenz-Punkte seiner STFT-Darstellung erfasst wird. Die Dünnbesetztheit der Sprachsignale im Zeit-Frequenz-Bereich eröffnet somit vielversprechende Möglichkeiten für eine bessere Störsignaldämpfung, wofür die Wahrscheinlichkeit berechnet werden muss, dass in einem Zeit-Frequenz-Punkt das Sprachsignal vorhanden ist. Diese Aufgabe wird vom vierten Baustein der Entstörung spektraler Amplituden übernommen, der als die SPP-Berechnung bezeichnet wird.

Um die Problemstellung der SPP-Berechnung statistisch zu formulieren, wird für jeden Zeit-Frequenz-Punkt eine binäre Zufallsvariable $H(k,\ell)$ mit zwei möglichen Realisierungen (oder Hypothesen) $\{H_0,\, H_1\}$ eingeführt, wobei $H(k,\ell) = H_0$ für die Abwesenheit des Sprachsignals im (k,ℓ)-ten Zeit-Frequenz-Punkt steht und $H(k,\ell) = H_1$ für die Sprachsignalpräsenz [SKS99]. Dabei ist die Zufallsvariable $H(k,\ell)$ mit der Beobachtung $Y(k,\ell)$ wie folgt verbunden:

$$H(k,\ell) = H_0: \quad Y(k,\ell) = D(k,\ell) \tag{2.29}$$
$$H(k,\ell) = H_1: \quad Y(k,\ell) = S(k,\ell) + D(k,\ell) \tag{2.30}$$

mit den bedingten Verteilungsdichtefunktionen der Beobachtung $Y(k,\ell)$ laut [Coh01]

$$p_{Y(k,\ell)|H_0}(y) = \mathcal{N}_{\mathbb{C}}(y;\lambda_D(k,\ell)). \tag{2.31}$$
$$p_{Y(k,\ell)|H_1}(y) = \mathcal{N}_{\mathbb{C}}(y;\lambda_S(k,\ell) + \lambda_D(k,\ell)) \tag{2.32}$$

Wird das *a posteriori* SNR $\gamma(k,\ell)$ als eine Beobachtung betrachtet, können mit einer Zufallsvariablentransformation die Verteilungsdichtefunktionen $p_{\gamma(k,\ell)|H_0}(\gamma)$ und $p_{\gamma(k,\ell)|H_1}(\gamma)$ berechnet werden, die sich als Exponentialverteilungen darstellen [Coh03, VM06]. Die Verwendung von $\gamma(k,\ell)$ statt $Y(k,\ell)$ trägt zur nummerischen Robustheit eines SPP-Schätzers bei, der in diesem Fall auf einer relativen Größe und nicht auf einer absoluten arbeitet. Dabei bleibt die statistische Modellierung für eine analytische Rechnung handhabbar und führt zur folgenden Erwartungswerten der beteiligten Zufallsvariablen $\gamma(k,\ell)|H_1$ und $\gamma(k,\ell)|H_0$:

$$\mathbb{E}[\gamma(k,\ell)\,|\,H_0] = 1, \tag{2.33}$$
$$\mathbb{E}[\gamma(k,\ell)\,|\,H_1] = \xi(k,\ell) + 1. \tag{2.34}$$

Die statistische Modellierung im γ-Bereich ermöglicht außerdem eine SPP-Berechnung, welche die Korrelationen zwischen benachbarten Zeit-Frequenz-Punkten der STFT-Koeffizienten des Sprachsignals geschickt ausnutzt [Coh01, GBM08].

Mit dieser statistischen Modellierung kann die gewünschte Sprachpräsenzwahrscheinlichkeit über die Bayessche-Regel berechnet werden:

$$\mathcal{P}(k,\ell) \;\triangleq\; \Pr(H(k,\ell) = H_1\,|\,\gamma(k,\ell)) \;=\; \frac{\Lambda(k,\ell)}{\Lambda(k,\ell)+1}, \tag{2.35}$$

wobei $\Lambda(k,\ell)$ sich als das generalisierte Likelihood-Verhältnis[7] (engl. *generalized Likelihood ratio*, GLR) darstellt, das wie folgt definiert ist:

$$\Lambda(k,\ell) \triangleq \frac{\Pr(H_1)}{\Pr(H_0)} \cdot \frac{p_{\gamma(k,\ell)|H_1}(\gamma)}{p_{\gamma(k,\ell)|H_0}(\gamma)} \tag{2.36}$$

mit den *a priori* Wahrscheinlichkeiten der Sprachsignalpräsenz $\Pr(H_1)$ und der Sprachsignalabwesenheit (engl. *speech absence probability*, SAP)

$$q \triangleq \Pr(H_0) = 1 - \Pr(H_1). \tag{2.37}$$

Während die *a priori* SAP q in vielen Verfahren wie [MM80, EM84, SKS99, SKY99, KC00, GBM08, FF12, MHA14, FG14, TVHU13, Gla15, MA$^+$16] als konstant (oder fixiert) angenommen wird, wird sie in den anderen wie [MCA99, MWJ00, Coh01, DA09] geschätzt. In der Literatur sind folgende konstante Werte der *a priori* SAP zu finden: $\Pr(H_0) = 0.2$ in [EM84, FF12], $\Pr(H_0) = \frac{1}{3}$ in [SKS99], $\Pr(H_0) = 0.4$ in [MA$^+$16], $\Pr(H_0) = 0.5$ in [MM80, Yan93, GBM08, FG14], $\Pr(H_0) = 0.6$ in [MHA14] oder $\Pr(H_0) \approx 0.94$ in [KC00].

Je nach Art der verwendeten Filterfunktion kann $\mathcal{P}(k,\ell)$ auf zwei unterschiedlichen Weisen in die Berechnung der resultierenden Filterfunktion einfließen, entweder multiplikativ oder als eine Potenz. Die multiplikative Variante wie in [MM80, EM84, Yan93] resultiert in folgender Filterfunktion zur Entstörung spektraler Amplituden:

$$G(k,\ell) = \mathcal{P}(k,\ell) \cdot G_{H_1}(k,\ell) + (1 - \mathcal{P}(k,\ell)) \cdot G_{H_0}, \tag{2.38}$$

wobei $G_{H_1}(k,\ell)$ die Filterfunktion in der Anwesenheit des Sprachsignals ist und G_{H_0} in seiner Abwesenheit. Während als $G_{H_1}(k,\ell)$ eine beliebige Filterfunktion der Entstörung spektraler Amplituden wie z. B. das Wiener-Filter (2.24) verwendet wird, kann G_{min} für G_{H_0} eingesetzt werden, um das musikalische Rauschen klein zu halten. Allerdings ist auch die Anwendung von $G_{H_0} = 0$ denkbar [MM80, EM84, MCA99].

Verwendet man eine Filterfunktion für Entstörung logarithmischer spektraler Amplituden, wie z. B. die MMSE-LSA-Filterfunktion (2.27), lässt sich die berechnete SPP über den Exponenten in die Filterfunktion integrieren [Coh01]

$$G(k,\ell) = G_{H_1}(k,\ell)^{\mathcal{P}(k,\ell)} \cdot G_{H_0}^{1-\mathcal{P}(k,\ell)}. \tag{2.39}$$

Verwendet man die MMSE-LSA-Filterfunktion (2.27) als $G_{H_1}(k,\ell)$ bezeichnet man die Filterfunktion in (2.39) auch als eine *optimally modified log-spectral amplitude* (OMLSA) Filterfunktion [CB01b]. Allerdings existiert auch eine *multiplicatively modified* LSA-Filterfunktion, die in [MCA99] mit $G(k,\ell) = \mathcal{P}(k,\ell) \cdot G_{\text{LSA}}(k,\ell)$ angegeben wird. Sowohl (2.38) als auch (2.39) gewährleisten die Erfüllung der Bedingung $G(k,\ell) \geq G_{H_0}$. Bei Bedarf kann auch $G(k,\ell)$ nach oben mit G_{max} mit Hilfe von (2.20) begrenzt werden.

Da die SPP $\mathcal{P}(k,\ell)$ laut (2.38) und (2.39) die Entstörungsregel direkt beeinflusst, können Schätzfehler in der berechneten *a posteriori* SPP zum Teil gravierende Folgen für spektrale Entstörung haben. Eine falsche Klassifizierung, etwa das Sprachsignal als das Störsignal einzustufen, soll möglichst vermieden werden, denn diese führt zur unerwünschten Unterdrückung von Sprachsignalkomponenten und somit zur Verzerrung des Sprachsignals [CG08]. Wird fälschlicherweise das Störsignal als das Sprachsignal deklariert, lässt

[7] Der Begriff *generalized* bezieht sich dabei auf die Verwendung des Vorfaktors $\Pr(H_1)/\Pr(H_0)$, ohne welchen $\Lambda(k,\ell)$ aus (2.36) zum konventionellen Likelihood-Verhältnis wird.

das System zur Signalentstörung das Störsignal ungedämpft durch. Somit gelangt zwar das Störsignal in das prozessierte Signal, aber das ist keine gravierende Verzerrung zumindest hinsichtlich der Qualität des entstörten Sprachsignals. Allerdings leidet in diesem Fall die Störsignaldämpfung. Aus diesen Gründen ist eine robuste und möglichst fehlerfreie SPP-Berechnung notwendig.

2.3. Bewertungsmaße der spektralen Sprachsignalentstörung

Die Leistungsfähigkeit des Gesamtsystems zur spektralen Entstörung wie in Abb. 2.1 kann auf unterschiedlichen Wegen bewertet werden. Zum einen gibt es die Möglichkeit, einen subjektiven Eindruck beim Anhören des prozessierten Sprachsignals $\hat{s}(n)$ zu bekommen. Zum anderen können die objektiven Maße berechnet werden, welche die Verständlichkeit, die Qualität oder die durch das System hervorgerufene Störsignaldämpfung messen, wofür allerdings das ungestörte Sprachsignal $s(n)$ vorliegen muss [HL08]. Ist man an der Bewertung der einzelnen Bausteine der spektralen Entstörung aus Abb. 2.2 interessiert, kann man objektive Fehlermaße definieren, welche z. B. den mittleren Schätzfehler oder die mittlere Schätzfehlervarianz messen. Dafür muss wiederum eine Referenz gegeben sein, welche sich als Sollwert der zu schätzenden Größe darstellt. In diesem Abschnitt werden kurz die objektiven Maße beschrieben, die in den Untersuchungen dieser Arbeit häufig verwendet werden.

Messung der Verständlichkeit bei Sprachsignalen: Um Verständlichkeit der Sprachsignale objektiv zu messen, wird das *short-time objective intelligibility* (STOI) Maß verwendet [THHJ10]. Bei Messung der Verständlichkeit der im Zeit-Frequenzbereich entstörten Sprachsignale korrelierte das STOI-Maß sehr stark mit prozentualer Anzahl der korrekt erkannten Wörter, die in den Experimenten mit den Testpersonen gemessen wurde. In den Untersuchungen in [THHJ11] nahm der entsprechende Korrelationskoeffizient die Werte höher als 0.92 an. Das STOI-Maß ist im Wertebereich $[0,\, 1]$ definiert, wobei höhere Werte bessere Verständlichkeit bedeuten. Die STOI-Werte können auch in Prozent $\mathrm{STOI}\,/\,\% \;=\; 100 \cdot \mathrm{STOI}$ angegeben werden. Die Verbesserung der Sprachsignalverständlichkeit, welche durch die Verarbeitung in einem Entstörungssystem erreicht wird, kann mit Hilfe von $\Delta\mathrm{STOI} = \mathrm{STOI}_{\mathrm{OUT}} - \mathrm{STOI}_{\mathrm{IN}}$ gemessen werden[8], die sich als Differenz von STOI-Werten am Ausgang des Systems $\mathrm{STOI}_{\mathrm{OUT}}$ und an seinem Eingang $\mathrm{STOI}_{\mathrm{IN}}$ darstellt. Eine gute Zusammenstellung einiger weiteren Messgrößen für eine objektive Messung von Verständlichkeit der Sprachsignale ist in [MHL09] zu finden.

Bewertung der Sprachsignalqualität: Für die Messung der Qualität der Sprachsignale werden die von *International Telecommunication Union* (ITU) empfohlenen Maße verwendet [ITU07]. Davon gibt es drei unterschiedlichen objektiven Messgrößen, welche die subjektiven *mean opinion score* (MOS) Werte nachbilden, die in [ITU96] zur Beurteilung der Sprachsignalqualität in den Telekommunikationsdiensten eingeführt werden und Werte zwischen 1 und 5 annehmen können. Während ein MOS-Wert von 1 eine sehr schlechte Signalqualität bedeutet, entspricht der MOS-Wert von 5 einer hervorragenden Qualität. Die erste objektive Messgröße ist das *perceptual evaluation of speech quality* (PESQ) Maß, das eine Ähnlichkeit zwischen $\hat{s}(n)$ und $s(n)$ misst [ITU01]. Obwohl der Definition nach das

[8] Die Abkürzungen IN und OUT stehen dabei jeweils für Eingang (*input*) und Ausgang (*output*).

PESQ-Maß im Wertebereich $(-0.5;\, 4.5)$ liegt, sind die PESQ-Werte kleiner 1 sehr selten, was einen Vergleich mit subjektiven MOS-Werten etwas vereinfacht. Für einen direkten Vergleich der PESQ-Werte mit den MOS-Werten wird allerdings noch ein monotones Polynom dritten Grades benötigt, das in [ITU01] leider nicht weiter spezifiziert wird.

Um einen direkten (linearen) Vergleich der PESQ-Werte mit den MOS-Werte zu ermöglichen, wird in [ITU03] eine Abbildungsfunktion standardisiert. Mit dieser werden die PESQ-Werte in die *mean opinion score - listening quality objective* (MOS-LQO) Werte umgerechnet, die der Definition nach im Wertebereich $(1.02;\, 4.56)$ liegen[9]. Da das PESQ-Maß aus [ITU01] und das MOS-LQO-Maß aus [ITU03] für die Messung der Sprachsignalqualität in den schmalbandige Telekommunikationssystemen entwickelt wurden, sind sie für die Anwendung im Frequenzbereich zwischen $300\,\text{Hz}$ und $3.1\,\text{kHz}$ konzipiert. Für die Messung der Qualität von breitbandigen Signalen wurde in [ITU07] das breitbandige (engl. *wide-band*, WB) MOS-LQO ($\text{MOS-LQO}_{\text{WB}}$) Maß standardisiert, das für den Frequenzbereich zwischen $50\,\text{Hz}$ und $7\,\text{kHz}$ ausgelegt ist. Um Missverständnisse zu vermeiden, werden die in [ITU01] und in [ITU03] standardisierten Qualitätsmaße mit dem Präfix schmalbandig (engl. *narrow-band*, NB) versehen und im Weiteren als PESQ_{NB} und $\text{MOS-LQO}_{\text{NB}}$ bezeichnet. Man beachte, dass die Qualität der mit $16\,\text{kHz}$ abgetasteten Sprachsignale durchaus mit den Maßen PESQ_{NB} und $\text{MOS-LQO}_{\text{NB}}$ gemessen werden kann [ITU07]. Ein großer Vorteil der Verwendung der durch ITU-standardisierten Qualitätsmaße ist die Bereitstellung einer *American National Standards Institutes* (ANSI) Realisierung zur Berechnung von PESQ_{NB}, $\text{MOS-LQO}_{\text{NB}}$ und $\text{MOS-LQO}_{\text{WB}}$ in Programmiersprache C, die für wissenschaftliche Zwecke frei zugänglich ist. Ein weiteres Argument für die Verwendung dieser Qualitätsmaße ist ihre hohe Korrelation mit subjektiven MOS-Werten, wie Untersuchungen mit Testpersonen gezeigt haben [HL08].

Messung der Störsignaldämpfung: Um quantitativ messen zu können, wie stark das Störsignal im Rahmen der Sprachsignalentstörung unterdrückt wird, wird ähnlich wie in [YF11] das globale SNR verwendet, das sowohl am Eingang SNR_{IN} als auch am Ausgang SNR_{OUT} eines Systems aus Abb. 2.1 berechnet wird. Unter der Annahme, dass neben dem gestörten Sprachsignal $y(n)$ auch das ungestörte Signal $s(n)$ und somit auch das Störsignal $d(n)$ für die Systembewertung vorliegen, kann das SNR_{IN} im Zeitbereich folgendermaßen berechnet werden:

$$\text{SNR}_{\text{IN}}\,/\text{dB} = 10 \cdot \log_{10} \frac{\sum_{n=1}^{N} s^2(n)}{\sum_{n=1}^{N} d^2(n)}, \tag{2.40}$$

wobei N die gesamte Anzahl der Abtastwerte in den Signalen $s(n)$ und $d(n)$ ist, die mit der Signallänge einhergeht. Außerdem können Signale $s(n)$ und $d(n)$ jeder für sich mit Hilfe der spektralen Filterfunktion $G(k,\ell)$ aus (2.17) gefiltert werden, die für die Entstörung der spektralen Amplituden $|Y(k,\ell)|$ des gestörten Sprachsignals aus (2.1) verwendet wurde. Daraus resultieren die gefilterten Signale $\tilde{s}(n)$ und $\tilde{d}(n)$, die sich aufgrund der Linearität des Signalmodels (2.1) und der STFT zum entstörten Sprachsignal $\hat{s}(n)$ aus Abb. 2.1 zusammensetzen, so dass es gilt $\hat{s}(n) = \tilde{s}(n) + \tilde{d}(n)$. Da die spektrale Funktion $G(k,\ell)$ für die Berechnung von $\tilde{s}(n)$ und $\tilde{d}(n)$ als gegeben angenommen wird, kann man von einer *white-box* Bewertungsmethode sprechen[10] [GMV96]. Aus den Signalen $\tilde{s}(n)$ und $\tilde{d}(n)$, die jeweils

[9] Wird die Qualität der Sprachsignale in den Auswertungen mit Testpersonen gemessen, bezeichnet man das resultierende Bewertungsmaß als *mean opinion score - listening quality subjective* (MOS-LQS) [ITU07]. Somit kann MOS-LQO als eine Schätzung von MOS-LQS betrachtet werden.

[10] Ist ein Zugriff auf die internen Systemgrößen für ihre Verwendung in der Systembewertung nicht gegeben,

als *white-box* Komponenten des ungestörten Sprachsignals und des Störsignals bezeichnet werden, wird nun das globale ausgangsseitige SNR (SNR_{OUT}) berechnet

$$\text{SNR}_{\text{OUT}}\,/\text{dB} = 10 \cdot \log_{10} \frac{\sum_{n=1}^{N} \tilde{s}^2(n)}{\sum_{n=1}^{N} \tilde{d}^2(n)}. \tag{2.41}$$

Für die Bewertung der Leistungsfähigkeit eines Systems hinsichtlich der Störsignaldämpfung kann anschließend wie in [YF11] eine durch das System hervorgerufene Verbesserung des globalen SNR-Wertes verwendet werden:

$$\Delta\text{SNR}\,/\text{dB} = \text{SNR}_{\text{OUT}} - \text{SNR}_{\text{IN}}. \tag{2.42}$$

Das Konzept der Δ Größen kann auch auf alle anderen Maße der Verständlichkeit und der Signalqualität angewandt werden. Der Vollständigkeit halber soll an dieser Stelle erwähnt werden, dass ein segmentelles SNR in der Bewertung der Störsignaldämpfung oft verwendet wird [QBC88, CG08].

Sollen die einzelnen Bausteine eines Systems zur spektralen Entstörung in Abb. 2.2 objektiv bewertet werden, greift man zur anderen Bewertungsmaßen. Da die Bausteine 2, 3 und 4 sich direkt am Systemausgang befinden, werden sie in den Anwendungen zur spektralen Sprachsignalentstörung häufig mit den bereits beschriebenen Bewertungsmaßen begutachtet. Etwas anders sieht die Bewertung der RLDS-Schätzer, auf die im Weiteren etwas ausführlicher eingegangen werden soll.

Bewertung der RLDS-Schätzverfahren: Die Bewertungsmaße der RLDS-Schätzer werden unter der Annahme berechnet, dass für die Bewertung sowohl die Schätzwerte $\hat{\lambda}_D(k,\ell)$ als auch die Referenz $\lambda_D(k,\ell)$ vorliegen. Als Maß für den mittleren Schätzfehler kann ein mittlerer logarithmischer Schätzfehler (engl. *log-error mean*, LEM) aus [HJH08, HHJ10] verwendet werden:

$$\text{LEM} = \frac{1}{K \cdot L} \sum_{k=1}^{K} \sum_{\ell=1}^{L} \Delta(k,\ell), \tag{2.43}$$

mit dem logarithmischen Schätzfehler im (k,ℓ)-ten Zeit-Frequenz-Punkt

$$\Delta(k,\ell) = \left| 10 \log_{10} \frac{\lambda_D(k,\ell)}{\hat{\lambda}_D(k,\ell)} \right|. \tag{2.44}$$

Man beachte, dass das LEM-Maß in [HJH08] als ein symmetrisches Fehlermaß eingeführt wird, das sowohl Überschätzung als auch Unterschätzung im gleichen Maße bestraft. Das LEM-Maß wird außerdem in einigen weiteren Publikationen wie [EH08, Yu09, HHJK09] verwendet. In [GH11] und [GH12] wird es zusätzlich in zwei getrennte Anteile für eine Überschätzung und einer Unterschätzung aufgeteilt, die eine genauere Analyse des Schätzfehlers erlauben. Die Verwendung von logarithmischen spektralen Größen ist weit verbreitet in den Messgrößen in der Sprachsignalverarbeitung [GBGM80, SA05, IS06]. Allerdings gibt es auch die nicht logarithmischen Fehlermaße, die bei der Bewertung der RLDS-Schätzung eingesetzt werden können [Mar01, Coh03, YS05, KSS06, RL06].

können die sogenannten *black-box* Bewertungsmethoden angewandt werden [FSS08].

Als ein Maß für die Schätzfehlervarianz wird die mittlere Varianz des logarithmischen Schätzfehlers (engl. *log-error variance*, LEV) verwendet, die in [TTM+11] als *variance of the logarithmic difference* bezeichnet wird. Die LEV ist folgendermaßen definiert:

$$\text{LEV} = \frac{1}{V \cdot M} \sum_{v=1}^{V} \sum_{m=1}^{M} \text{LEV}(v, m) \tag{2.45}$$

mit der Varianz des logarithmischen Schätzfehlers $\text{LEV}(v, m)$ im (v, m)-ten Unterblock

$$\text{LEV}(v, m) = \frac{1}{K_s \cdot L_s} \sum_{k=vK_s+1}^{(v+1)K_s} \sum_{\ell=mL_s+1}^{(m+1)L_s} \left(\Delta(k, \ell) - \bar{\Delta}(v, \ell)\right)^2 \tag{2.46}$$

und dem mittleren logarithmischen Schätzfehler im v-ten Unterblock und ℓ-ten Rahmen

$$\bar{\Delta}(v, \ell) = \frac{1}{K_s} \sum_{k=vK_s+1}^{(v+1)K_s} \Delta(k, \ell). \tag{2.47}$$

Dabei sind M und V jeweils die Anzahl der Unterblöcke entlang der Zeit und Frequenz, über welche die $\text{LEV}(v, m)$ gemittelt wird. Der Wert von $\text{LEV}(v, m)$ wird für jedes Zeit-Frequenz Unterblock ausgerechnet, der in dem Rahmen $m \cdot L_s + 1$ und in dem Frequenzband $v \cdot K_s + 1$ beginnt und die Länge $L_s = 2K_s$ Rahmen und K_s Frequenzbänder beinhaltet. Je größer der Wert von LEV ist, desto mehr Artefakte, wie z. B. *musical tones*, sind im prozessierten Signal vorhanden [TTM+11]. Alternativ könnte man für die Bewertung der Schätzfehlervarianz auch eine globale Varianz des logarithmischen Schätzfehlers $\Delta(k, \ell)$ wie in [Yu09] verwenden.

Obwohl bei der Bewertung der Schätzer das LEM-Maß höhere Priorität als LEV hat, ist die Berechnung von LEV von großer Bedeutung. Weisen zwei Schätzer ähnliche LEM-Werte auf, soll der Schätzer mit den kleineren LEV-Werten bevorzugt werden [TTM+11]. Man beachte, dass für die beiden Fehlermaße LEM und LEV gilt: kleinere Werte bedeuten bessere Leistungsfähigkeit des Schätzers. Der Vollständigkeit halber soll an dieser Stelle erwähnt werden, dass in den Untersuchungen dieser Arbeit neben den Maßen LEM und LEV auch andere Fehlermaße für die Bewertung der Schätzer verwendet werden, die in den jeweiligen Kapiteln auch kurz beschrieben werden.

Ein weiterer Punkt, der bei der Bewertung der RLDS-Schätzung betrachtet werden muss, ist die Wahl der Referenz $\lambda_D(k, \ell)$, von der unklar ist, wie sie in einer konkreten Anwendung berechnet werden soll, selbst wenn der aktuelle Wert der momentanen spektralen Leistung $|D(k, \ell)|^2$ eines Störsignals vorliegt [Mar01, GH11]. Eine Möglichkeit, die in der Literatur für die Berechnung der Referenz $\lambda_D(k, \ell)$ vorgeschlagen wird, ist die Glättung der aufeinander folgenden Werte $|D(k, \ell)|^2$ in jedem Frequenzband entlang der Zeitachse mit einem *infinite impulse response* (IIR) Filter erster Ordnung:

$$\lambda_D(k, \ell) = \alpha_\text{D} \cdot \lambda_D(k, \ell - 1) + (1 - \alpha_\text{D}) \cdot |D(k, \ell)|^2, \tag{2.48}$$

wobei $\alpha_\text{D} \in [0;\, 1]$ eine festgelegte Glättungskonstante ist. Diese nimmt häufig den Wert von $\alpha_D = 0.9$ wie in [Mar01,HJH08,EH08,HHJK09,HHJ10,TTM+11] oder von $\alpha_D = 0.95$ wie in [Coh03] an. Obwohl die Glättungskonstanten hier etwas unterschiedlich sind, stehen sie

eigentlich für die rekursive Glättung mit sehr ähnlichen Verfolgungsgeschwindigkeiten von jeweils $v_P \approx 28.6\,\text{dB/s}$ und $v_P \approx 27.8\,\text{dB/s}$, siehe Anhang A. Dabei muss man beachten, dass je näher die Werte der Glättungskonstante zu 1 sind, desto glatter ist der Verlauf der Referenzwerte, aber auch desto größer die Latenz, die durch die Glättung (2.48) hervorgerufen wird. Mit steigendem α sinkt die Verfolgungsgeschwindigkeit laut (A.10). Das Vorhandensein der Latenz macht allerdings die Referenz parteiisch, denn RLDS-Schätzer mit einer ähnlichen Latenz werden in diesem Fall bevorzugt[11]. Um dieses Bevorzugen zu vermeiden, kann zum einen die Berechnung der Fehlermaße angepasst werden, wie in Kap. 7 gezeigt wird. Zum anderen kann die Referenz antikausal geglättet somit verzögerungsfrei berechnet werden, wie dies in Abschnitt 7.3 der Fall ist. Oder man verwendet die Referenz $\lambda_D(k,\ell) = |D(k,\ell)|^2$, wie in Abschnitt 5.4 und in Abschnitt 6.4, was mit $\alpha_\text{D} = 0$ einhergeht. Diese Referenzwahl kommt in der Bewertung der RLDS-Schätzer durchaus vor [SA05, YS05, KSS06, GKR12].

Neben LEM- und LEV-Maßen werden im Rahmen dieser Arbeit noch zwei weitere objektive Fehlermaße zur Bewertung der RLDS-Schätzung verwendet: das *root mean square error* (RMSE) Maß aus Unterabschnitt 7.2.2 und das spektrale Distanzmaß (engl. *spectral distance measure*, SDM) aus Abschnitt 5.4, das in [IS06] zum ersten Mal eingeführt wurde. Man beachte, dass die objektiven Bewertungsmaße auch zur Optimierung der RLDS-Schätzer eingesetzt werden können, wie dies in Abschnitt 6.3 der Fall ist.

2.4. Datenbanken für experimentelle Untersuchungen

In den experimentellen Untersuchungen dieser Arbeit wurden die Audiosignale einiger renommierter Datenbanken wie TIMIT, NOISEX-92 und CHiME-3 verwendet, die in diesem Kapitel kurz beschrieben werden. Während die TIMIT-Datenbank die ungestörten Sprachsignale beinhaltet, werden die Störsignale aus der NOISEX-92-Datenbank genommen, welche mit den Störsignalen der RSG-10 und SPIB-Datenbanken verwandt sind. Vereinzelt kommen auch die Störsignale der *SOUND-IDEAS*-Datenbank zum Einsatz [Nim92]. In der CHiME-3-Datenbank sind sowohl die ungestörten Sprachsignale der WSJ0-Datenbank als auch weitere Störsignale zu finden.

Ungestörte Sprachsignale der TIMIT-Datenbank: Die (*Texas Instruments and Massachusetts Institute of Technology, TIMIT*) Datenbank war ursprünglich für die Aufgaben der automatischen Spracherkennung (engl. *automatic speech recognition*, ASR) entwickelt und enthält Aufnahmen von insgesamt 630 Sprechern der 8 Dialekte des amerikanischen Englisch, die jeweils 10 phonetisch reiche Sätze gesprochen haben [FDGM86]. Zu jedem gesprochenen Satz liegen vier Arten der Dateien vor: das akustische Sprachsignal, der in diesem Signal gesprochene Satz, sowie die zeitlich-gelabelte Transkriptionen von gesprochenen Wörtern und Phonemen. Somit ist die Datenbank für unterschiedliche akustisch-phonetische Untersuchungen sehr gut geeignet [LKS89]. Ausgezeichnet durch eine große Anzahl an Sprechern beinhalten Audioaufnahmen der TIMIT-Datenbank auch eine bemerkenswerte Anzahl von 2342 verschiedenen Sätzen, von denen zwei sogenannte Kalibrierungssätze von allen Teilnehmern gesprochen werden mussten[12]. Ein Satz der TIMIT-Datenbank beinhaltet

[11] Wie in Abschnitt 3.1 gezeigt wird, kommt die sogenannte ausgangsseitige Glättungstechnik häufig zum Einsatz in der RLDS-Schätzung. Diese Technik sorgt für eine bestimmte Latenz der RLDS-Schätzwerte.

[12] Die Kalibrierungssätze '*She had your dark suit in greasy wash water all year*' und '*Don't ask me to carry an*

durchschnittlich etwa 8.2 Wörter und dauerte dementsprechend nur einige wenigen Sekunden [Byr94]. Während der Anteil der männlichen Sprecher etwa $70\,\%$ beträgt, beläuft sich der Anteil der Sprecherinnen auf nur etwa $30\,\%$. Die Aufzeichnungen wurden unter sehr guten akustischen Bedingungen mit Hilfe eines Nahsprechmikrofons aufgenommen mit einem SNR von etwa $30\,\text{dB}$ [ZSG90]. Obwohl man die ursprünglichen Sprachsignale mit der Abtastrate von $20\,\text{kHz}$ aufgezeichnet hat, wurden sie für die Transkription auf die Abtastrate von $16\,\text{kHz}$ umgesetzt und auch so anschließend auf einer CD-ROM (engl. *compact disc read-only memory*) veröffentlicht, die in [GLF$^+$93] ausführlich beschrieben wird. Obwohl die Signale der TIMIT-Datenbank in einen Trainingsset und einen Testset aufgeteilt sind, werden im Rahmen dieser Arbeit nur die Trainingsdaten verwendet.

Störsignale der RSG-10-, NOISEX-92- und SPIB-Datenbanken: Die Störsignale der weit verbreiteten NOISEX-92-Datenbank haben ihren Ursprung in der weniger bekannten RSG-10-Datenbank, die am holländischen *TNO Institute of Perception* von einer Forschungsgruppe (engl. *Research Study Group*, RSG) für Sprachsignalverarbeitung zusammengestellt wurde [SG88]. Die RSG-10-Datenbank beinhaltet insgesamt Störsignale von 24 unterschiedlichen Störarten, die sowohl künstliche stationäre Signale als auch hochgradig nichtstationäre Signale repräsentieren. Die letzten Signalaufnahmen wurden während bestimmter Militäraktionen aufgezeichnet. Von den Störsignalen der RSG-10-Datenbank wurden Signale von 8 Störarten für die Verwendung in der NOISEX-92-Datenbank ausgewählt, die für ASR Aufgaben in gestörten Umgebungen entwickelt wurde [VS93]. Bei der Vermarktung der NOISEX-92-Datenbank wurden allerdings alle Störsignale der RSG-10-Datenbank mitgeliefert, die man dann als Störsignale der NOISEX-92-Datenbank bezeichnete. Gleichzeitig startete an der *Rice University* ein Projekt zur Erstellung einer öffentlich zugänglichen kostenlosen Datenbanken für die Sprachsignalverarbeitung mit dem Namen *Signal Processing Information Base* (SPIB) [JS93]. Im Rahmen dieses Projektes entstand unter anderem die SPIB-Datenbank für Störsignale, welche folgende 15 Störarten der RSG-10-Datenbank repräsentierten: *babble, buccaneer1, buccaneer2, destroyerengine, destroyerops, f16, factory1, factory2, hfchannel, leopard, m109, machinegun, pink, volvo, white*[13]. Diese Signale, die man auch häufig als NOISEX-92 Daten bezeichnet, werden in den experimentellen Untersuchungen dieser Arbeit verwendet. Da die Signale ursprünglich mit der Abtastrate von $19.98\,\text{kHz}$ aufgezeichnet wurden, werden sie für die Verwendung gemeinsam mit den ungestörten Sprachsignalen der TIMIT-Datenbank auf die Abtastrate von $16\,\text{kHz}$ umgesetzt. Im Unterschied zu den relativ kurzen Signalen der TIMIT-Datenbank sind alle Störsignale der SPIB-Datenbank $235\,\text{s}$ lang.

Um das Verhalten der Systeme zur Sprachsignalentstörung in Umgebungen mit unterschiedlicher Störbelastung zu untersuchen, sollen die ungestörten Sprachsignale der TIMIT-Datenbank $s(n)$ und die Störsignale der SPIB-Datenbank $d(n)$ entsprechend (2.1) additiv überlagert werden. Da viele Verfahren zur RLDS-Schätzung (wie z. B. in [Mar94, Mar01]) eine gewisse Anlaufphase benötigen, bevor sie imstande sind, reguläre Schätzwerte berechnen zu können, werden kurze Aufnahmen der TIMIT-Datenbank zur längeren Sprachsignalen von eins, zwei oder drei Minuten zusammengesetzt. Somit befinden sich die RLDS-Schätzer zu Beginn einer neuen Äußerung bereits im eingeschwungenen Zustand (die aller-

oily rag like that' kommen in der Datenbank sehr häufig vor und weisen durchschnittliche Dauer von etwa $2.7\,\text{s}$ auf [Byr94]. Die Untersuchungen dieser Sätze zeigten, dass Männer schneller als Frauen sprachen.

[13] Zum Zeitpunkt der Verfassung dieser Arbeit waren die erwähnten Störsignale der SPIB-Datenbank immer noch frei zugänglich und zwar unter http://spib.linse.ufsc.br/noise.html.

erste Äußerung ausgenommen). Um die Anteile der männlichen und der weiblichen Sprecher in den Sprachdaten anzugleichen, werden die langen Sprachsignale jeweils für weibliche und männliche Sprecher generiert. Anschließend werden die gleich langen Signale $s(n)$ und $d(n)$ entsprechend dem globalen eingangsseitigen SNR zu den gestörten Sprachsignalen $y(n)$ aufaddiert. Ein Vorteil der Erzeugung von solchen künstlich generierten gestörten Sprachsignalen ist die Möglichkeit, die zu entwickelnden Schätzer unter kontrollierten Bedingungen zu untersuchen und bei Bedarf zu optimieren. Allerdings, unabhängig von der Vielfalt der künstlich generierten Daten, gibt es immer noch keine hundertprozentige Garantie, dass die Verfahren auf den Daten, die sie in der Entwicklungsphase nicht gesehen haben, genauso gut funktionieren.

Signale der CHiME-3-Datenbank: Von den Daten der dritten (engl. *computational hearing in multisource environments*, *CHiME*) Datenbank werden im Rahmen dieser Arbeit nur die Daten des sogenannten *isolated simulated development* Datensatzes verwendet, die in vier verschiedenen Störumgebungen vorliegen: in einem Bus (*bus*) in einer Cafeteria (*caf*), in einer Fußgängerzone (*ped*) und auf einer Straße (*str*) [BMVW15]. Die Tatsache, dass hier neben den gestörten Sprachsignalen $y(n)$ auch Signale $s(n)$ und $d(n)$ separat zu finden sind, ermöglicht die erwünschte Bewertung der Leistungsfähigkeit eines Systems zur Sprachsignalentstörung. Die Sprachaufforderungen dieser Daten stammen aus der (engl. *Wall Street Journal*, WSJ0) Datenbank [GGPP07]. Von jedem der vier Sprecher (zwei weiblich und zwei männlich) liegen jeweils 410 unterschiedliche kurze Äußerungen vor. Da die gesamte Dauer der Sprachmaterialien $2.88\,\mathrm{h}$ beträgt, liegt die durchschnittliche Dauer einer Äußerung bei etwa $6.32\,\mathrm{s}$. Die Daten wurden mit Hilfe eines Tablets mit 6 Mikrophonen aufgezeichnet, von denen die Signale des $5.$ Mikrophons für eine einkanalige Sprachsignalverarbeitung ausgewählt wurden. Alle Signale sind bereits mit $16\,\mathrm{kHz}$ abgetastet und weisen einen durchschnittlichen globalen SNR von etwa $5.8\,\mathrm{dB}$ und einen breitbandigen MOS-LQO$_{\mathrm{WB}}$-Wert von 1.27 auf. Somit sind die CHiME-3-Daten stark verrauscht und dementsprechend herausfordernd für eine Entstörung in einem System.

Bei einer Verarbeitung von kurzen Sprachsignalen der CHiME-3-Datenbank kann die Anlaufphase der verwendeten RLDS-Schätzer dadurch gewährleistet werden, dass der Anfang jeder Äußerung (z. B. der Dauer von $1\,\mathrm{s}$) vor dem Beginn dieser Äußerung rückwärts abgespielt wird und so einem RLDS-Schätzer präsentiert wird. Die auf diesem künstlich angehängten Signalstück berechneten RLDS-Schätzwerte werden anschließend einfach verworfen und fließen somit nicht in die Bewertung der Leistungsfähigkeit ein. Man beachte, dass diese Manipulation der Signale nur möglich ist, weil die zu verarbeitenden Sprachsignale aus den Datenbanken stammen und bereits vor der Signalverarbeitung in voller Länge im Systemspeicher vorliegen[14].

[14] Sonst, beim Einsatz auf einem Gerät zur Sprachsignalverarbeitung (wie z. B. einem Hörgerät) gibt es eine kurze Anlaufphase nur beim Einschalten des Gerätes, was der Verarbeitung von langen Signalen einer Datenbank entsprechen würde, wie dies bei den zusammengesetzten TIMIT-Daten der Fall ist.

3. Stand der Forschung

Nachdem die Grundlagen für Systeme zur spektralen Sprachsignalentstörung in Kap. 2 gelegt wurden, werden in diesem Kapitel verschiedene Verfahren erläutert, wie die Bausteine solcher Systeme konkret realisiert werden. Von einer bunten Palette der modernen RLDS-Schätzer werden in Abschnitt 3.1 zehn ausgewählte Schätzverfahren detailliert beschrieben. Dabei kristallisieren sich fünf verschiedene Grundtechniken heraus, die in der RLDS-Schätzung verfahrensübergreifend verwendet werden. In Abschnitt 3.2 werden anschließend spektrale Filterfunktionen der generalisierten Sprachsignalentstörung vorgestellt, die entweder mit Hilfe von generalisierten spektralen Amplituden hergeleitet werden oder auf einer statistischen Modellierung mit den generalisierten Verteilungsdichten basieren. Da das *Decision-Directed* Verfahren als ein *a priori* SNR-Schätzer in der überwiegenden Mehrheit von Systemen eingesetzt wird, wird es in Abschnitt 3.3 ausführlich vorgestellt. Neben seinen zahlreichen Modifikationen werden hier auch einige alternative Verfahren zur *a priori* SNR-Schätzung präsentiert. Im nächsten Abschnitt 3.4 werden verschiedene Verfahren zur Berechnung der Sprachpräsenzwahrscheinlichkeit beschrieben. Dabei wird der Schwerpunkt auf die Verfahren gelegt, welche die Korrelationen zwischen den benachbarten Zeit-Frequenz-Punkten berücksichtigen, die in den Sprachsignalen zwangsläufig vorhanden sind. Zum Schluss in Abschnitt 3.5 wird auf den Einsatz von tiefen neuronalen Netzen eingegangen, die in letzter Zeit zunehmend für spektrale Sprachsignalentstörung eingesetzt werden.

3.1. Schätzung der spektralen Rauschleistungsdichte

Groß ist die Vielfalt der ausgeklügelten Verfahren zur zuverlässigen Schätzung der spektralen Rauschleistungsdichte aus den einkanaligen Audioaufnahmen. Verschieden sind auch ihre Ansätze, um das große Ziel zu erreichen: eine robuste RLDS-Schätzung in der Gegenwart einer akustischen, in der Regel nichtstationären Störung. Diese Aufgabe ist sehr anspruchsvoll, denn eine der grundlegendsten Annahmen, die bei der Herleitung der Rauschschätzer gemacht wird, ist die Stationarität (oder Quasi-Stationarität) der Störung. Und da diese Annahme von den nichtstationären Störungen eindeutig verletzt wird, geht die Suche nach den effizienten und leistungsfähigen RLDS-Schätzern weiter.

Weist eine nichtstationäre Störung eine besondere Eigenschaft auf, die sie vom Sprachsignal unterscheidet, wird diese Eigenschaft gerne dazu genutzt, die Störung im gestörten Signal zu finden, um sie anschließend aus dem Signal zu entfernen. Ein Beispiel dafür sind die sogenannten transienten Störungen, die kurzzeitig und breitbandig sind wie Tastaturgeklapper, Geräusche einer tickenden Uhr, Türklopfen u. a.. Solche Störungen treten lokal im Spektrogramm des gestörten Signals auf und haben eine ihnen eigene Dynamik, die sie von einem Sprachsignal unterscheiden lässt. In [HDTC12] wird z. B. die Kurzzeitigkeit der tran-

sienten Ereignisse verbunden mit einem plötzlichen Anstieg und anschließendem Abfall des Signalpegels dafür genutzt, die Rauschleistungsdichte dieser Ereignisse zu schätzen. Dafür wird ein separater Rauschschätzer entwickelt, der nur darauf ausgerichtet ist, das LDS der transienten Ereignisse im gestörten Sprachsignal zu bestimmen. In [SMP13] wird sogar die Phase des gestörten Signals analysiert, um transiente Störungen zu finden, denn ein transientes Ereignis ist im Zeitbereich impulsartig und erzwingt somit im Signalspektrum kurzzeitig eine linear ansteigende Phase. In [TCG13] wird der Wiederholungscharakter der transienten Störungen ausgenutzt, um das LDS des abrupten Teils dieser Ereignisse im Rahmen einer sogenannten nicht-lokalen Filterung besser zu schätzen.

Die nichtstationären Hintergrundstörungen weisen allerdings nicht immer solche besonderen Eigenschaften auf wie die transienten Ereignissen. Dabei kommen sie überall vor: in einem Cafe, auf der Straße, an einem Arbeitsplatz oder auch Zuhause. Sie haben ähnliche Dynamik wie ein Sprachsignal oder weisen vergleichbare Korrelationseigenschaften auf, und doch sind sie in einem Sprachsignal als Störung einzuordnen. Die RLDS-Schätzverfahren, die sich mit solchen Störungen beschäftigen, müssen entweder mit Hilfe eines internen SPP-Schätzers zuverlässig unterscheiden können, welche Zeit-Frequenz-Punkte vom Sprachsignal und welche von der Störung dominiert werden, oder sie müssen auf andere Techniken zurückgreifen, die ganz ohne SPP-Schätzung auskommen, die in einer Anwendung immer fehlerbehaftet ist.

3.1.1. Einige konventionelle Schätzverfahren

Eine Reihe von modernen Verfahren zur LDS-Schätzung der Hintergrundstörungen werden in [TTM+11] hinsichtlich ihrer Leistungsfähigkeit verglichen. Dabei kommen 8 unterschiedliche Rauschschätzer zum Einsatz [Mar01, CB01a, Coh03, FRB07, HJH08, KPC09, Yu09, HHJ10], die in Gegenwart unterschiedlicher Störungen ausgewertet werden. Die Studie untersucht diese Schätzer hinsichtlich des mittleren logarithmischen Schätzfehlers LEM und der mittleren Varianz des logarithmischen Schätzfehlers LEV. Zwei weitere erwähnenswerte RLDS-Schätzer sind in [Hir93, HE95] und [GH11, GH12] zu finden. Im Weiteren werden diese zehn RLDS-Schätzer in der chronologischen Reihenfolge genauer betrachtet.

1. VAD-RA: Das erste Verfahren zur RLDS-Schätzung wird in [Hir93] eingeführt und in [HE95] in einem ASR-System eingesetzt. Eine RLDS-Schätzung wird hier in jedem Frequenzband in zwei Schritten ausgerechnet. Im ersten Schritt wird im Rahmen einer *voice activity detection* (VAD) in einem Vergleich mit einer Schwelle eine einfache harte Entscheidung zwischen den binären Werten 0 oder 1 getroffen, ob der aktuelle Zeit-Frequenz-Punkt vom ungestörten Sprachsignal jeweils dominiert wird oder nicht[1]. Liegt ein VAD-Schätzwert gleich 0 vor, sodass $Y(k,\ell) \approx D(k,\ell)$ vermutet wird, wird das Spektrogramm des verrauschten Signals $|Y(k,\ell)|^2$ mit einer rekursiven Mittelung (engl. *recursive averaging*, RA) erster Ordnung wie in (2.47) mit einer festgelegten Glättungskonstante gefiltert und als RLDS-Schätzwert $\hat{\lambda}_D(k,\ell)$ ausgegeben. Resultiert die VAD-Entscheidung in 1, wird die zeitliche Glättung außer Kraft gesetzt und der RLDS-Schätzwert konstant gehalten $\hat{\lambda}_D(k,\ell) = \hat{\lambda}_D(k,\ell-1)$. Das Verfahren ist somit sehr recheneffizient, hat nur zwei Parameter (der eine fließt in die Berechnung des Schwellenwertes ein und der andere ist die Glättungskonstante) und wird im Weiteren als VAD-RA-Schätzer bezeichnet. Die Schwäche des VAD-

[1] Im Rahmen dieser Arbeit wird eine VAD im Zeit-Frequenz-Bereich definiert. Man beachte, dass eine VAD alternativ auch im Zeitbereich definiert werden kann.

RA-Schätzers ist seine direkte Abhängigkeit von der einfach ausgelegten VAD-Komponente, die in einer Realisierung zu fehlerbehafteten RLDS-Schätzwerten führen kann.

2. OSMS: Ein RLDS-Schätzer, der ohne VAD-Komponente auskommt, ist das *Minimum Statistics* (MS) Verfahren, dessen erste Realisierung in [Mar94] vorgestellt wird. Viel bekannter ist jedoch seine Weiterentwicklung aus [Mar01]. Die beiden Verfahren nutzen die Tatsache aus, dass das LDS des gestörten Sprachsignals regelmäßig auf das Niveau des LDS des Störsignals absinkt. Dies geschieht in den zahlreichen Zeit-Frequenz-Punkten ohne Sprachaktivität, die jedoch im Rahmen der beiden Verfahren nicht explizit geschätzt werden muss. Der aktuelle minimale Wert des LDS des Störsignals wird in jedem Frequenzband durch eine kausale Minimumsuche bestimmt, die über die geglätteten LDS-Werte des gestörten Signals der letzten Sekunde durchgeführt wird. Die Länge des kausalen Fensters für die Minimumsuche wird so gewählt, dass die maximale erwartete Dauer der Sprachaktivität von diesem Fenster überbrückt wird. Somit ist sichergestellt, dass im Suchfenster immer Zeit-Frequenz-Punkte vorhanden sind, die vom Störsignal dominiert werden. Die gesuchte mittlere Rauschleistungsdichte wird aus den resultierenden Minimalwerten durch die Biaskorrektur ausgerechnet, die von den statistischen Eigenschaften der Minimalwerte abhängt (daher rührt auch der Name des Verfahrens). Während in [Mar94] die vorgeschlagenen Parameter für die rekursive Glättung und für die Biaskorrektur zeitinvariant sind, wird in [Mar01] eine optimale Glättungskonstante und eine ausgeklügelte Biaskorrektur vorgeschlagen, die zu zeitvarianten Parametern führen. Von der Leistungsfähigkeit her wird das MS-Verfahren vom *optimally smoothed* MS (OSMS) Verfahren aus [Mar01] in den Schatten gedrängt. Das OSMS-Verfahren genießt daher große Popularität, wird jedoch häufig auch als MS-Verfahren bezeichnet wie in [HJH08], [Yu09] oder [HHJ10]. Das OSMS-Verfahren hat prinzipiell einen einzigen Parameter - die Länge der Fenster für die Minimumsuche[2]. Ein offensichtlicher Nachteil der beiden Verfahren besteht allerdings darin, dass sie nicht schnell genug auf einen steigenden Pegel des Störsignals reagieren, was in der kausalen Minimumsuche begründet ist. Außerdem ist zu erwähnen, dass die in [Mar01] empfohlene Glättung nur im MMSE-Sinne optimal ist. Diese Tatsache wird in Kap. 5 noch ausführlich diskutiert.

3. MCRA: Während die rekursive Glättung des OSMS-Verfahrens ausschließlich der Minimumsuche dient, wird sie im (engl. *minima controlled recursive averaging*, MCRA) Verfahren aus [CB01a] direkt für die Berechnung der resultierenden Schätzwerte der Rauschleistungsdichte eingesetzt, wie dies auch beim VAD-RA-Verfahren der Fall ist. Dabei wird die zeitvariante Glättungskonstante durch eine Vorstufe gesteuert, in der die Sprachpräsenzwahrscheinlichkeit geschätzt wird. Nimmt die SPP im aktuellen Zeit-Frequenz-Punkt den Wert 1 an, wird die Glättungskonstante auch auf den Wert 1 gesetzt und somit die Rauschleistungsdichte auf dem zuletzt vom MCRA-Verfahren geschätzten Wert gehalten. Sinkt die SPP, darf auch die Glättungskonstante kleinere Werte annehmen, jedoch nicht kleiner als eine vorgegebene Konstante von 0.95. Das bedeutet, die Verfolgungsgeschwindigkeit v_P dieser Glättung bleibt kleiner als etwa $27.8\,\text{dB/s}$. Die SPP des MCRA-Verfahrens wird in der Vorstufe mit den vier folgenden Schritten geschätzt. Im ersten Schritt werden zunächst die aktuellen Werte des LDS des gestörten Signals entlang der Frequenzachse mit einem kurzen *finite impulse response* (FIR) Filter und anschließend entlang der Zeitachse mit einer zeitinvarianten Glättungskonstante von $0{,}8$ rekursiv geglättet, was einer hohen Verfolgungsgeschwindigkeit

[2] Im Rahmen einer effizienten Realisierung der kausalen Minimumsuche, die in [Mar94, Mar01] vorgestellt wird, wird das N_{MS} lange Fenster für Minimumsuche in U_{MS} Unterfenster der Länge V_{MS} aufgeteilt. Je nach der Fensterlänge N_{MS} müssen beim OSMS-Verfahren beide Parameter U_{MS} und V_{MS} gewählt werden.

von etwa $v_P = 121\,\text{dB/s}$ entspricht. Die Glättung entlang der Frequenzachse hat zum Ziel, die Korrelationen eines Sprachsignals entlang der Frequenz auszunutzen. Im zweiten Schritt wird ähnlich wie beim OSMS-Verfahren eine Minimumsuche über die vergangenen Werte des geglätteten LDS des gestörten Signals durchgeführt. Im dritten Schritt wird zunächst das Verhältnis vom aktuellen Wert des geglätteten LDS des gestörten Signals zum Ergebnis der Minimumsuche ausgerechnet und anschließend mit einer vorgegebenen Schranke verglichen. Ist das ausgerechnete Verhältnis größer als die Schranke, wird ein Indikator der Sprachpräsenz auf 1 und sonst auf 0 gesetzt. Im letzten vierten Schritt werden die SPP-Werte über die rekursive Glättung der berechneten Indikatorwerte mit einer weiteren zeitinvarianten Glättungskonstante von $0{,}2$ berechnet, die mit einer sehr hohen Verfolgungsgeschwindigkeit von etwa $v_F = 1.75\,\text{dB/ms}$ einhergeht. Die vielen Berechnungsschritte sorgen für die relativ große Anzahl von insgesamt 6 Parametern des MCRA-Verfahrens, die heuristisch gewählt werden. Im Vergleich zum OSMS-Verfahren reagiert der MCRA-Rauschschätzer schneller auf den steigenden Rauschpegel, denn seine Minimumsuche beeinflusst nur die Steuerung der Glättungskonstante der ausgangsseitigen Glättung. Sonst dient der MCRA-Schätzer als ein Ausgangspunkt für die Entwicklung einiger weiteren MCRA-basierten Verfahren.

4. IMCRA: Der *improved* MCRA (IMCRA) Rauschschätzer aus [Coh03] ist eine Weiterentwicklung des MCRA-Verfahrens. Dabei wird das IMCRA-Verfahren im Vergleich zum MCRA-Verfahren wie folgt verbessert: die Minimumsuche während der Sprachaktivität, das Schätzen von SPP und die Reduzierung des mittleren Schätzfehlers [Coh03]. Dafür wird die Vorstufe zur Schätzung von SPP in zwei Iterationsschritten realisiert. In der ersten Iteration wird zunächst die grobe SPP berechnet, um die Zeit-Frequenz-Punkte zu finden, in denen das Sprachsignal mit hoher spektralen Leistungsdichte vorliegt. Diese Zeit-Frequenz-Punkte werden aus der Signalverarbeitung der Vorstufe in der zweiten Iteration herausgenommen, denn sie erschweren nur das Schätzen der Rauschleistungsdichte. Dadurch wird die Suche des aktuellen minimalen Wertes der Rauschleistungsdichte robuster, sodass man kürzere Fenster der Minimumsuche und kleinere Konstanten für die rekursive zeitliche Glättung verwenden kann, die vor der Minimumsuche ausgeführt wird. Somit verbessern sich auch die Verfolgungseigenschaften des Schätzers in der Anwesenheit des nichtstationären Rauschens. Die stabilere Minimumsuche der zweiten Iteration führt außerdem zur feineren SPP-Schätzung, die im [Coh03] basierend auf der statistischen Modellierung vorgeschlagen wird. Außerdem wird ein analytischer Ausdruck für die Berechnung des Biaskompensationsfaktors des Rauschschätzers gefunden, der mit den statistischen Eigenschaften des SPP-Schätzers der zweiten Iteration einhergeht. Die zusätzlichen Arbeitsschritte des IMCRA-Verfahrens führen zwar zur besseren Leistungsfähigkeit des Verfahrens im Vergleich zum MCRA-Schätzer, verdoppeln allerdings zugleich die Anzahl der einzustellenden Parameter und verkomplizieren somit seine Parametrisierung.

5. EMCRA: Da die robuste Minimumsuche mit der einfachen rekursiven Glättung im Rahmen der Schätzung der Rauschleistungsdichte elegant kombiniert werden kann, greifen auch weitere Entwickler dieses Konzept in ihren Forschungsarbeiten auf. Ein weiteres Beispiel dafür ist das *enhanced* MCRA (EMCRA) Verfahren aus [FRB07], das eine Antwort auf die Frage sucht: wie man das Fenster der Minimumsuche verkleinert, um dadurch die Verfolgungseigenschaft des Rauschschätzers im nichtstationären Rauschen zu verbessern, ohne dabei die Zeit-Frequenz-Punkte mit schwachem LDS des Sprachsignals fälschlicherweise als Punkte mit Sprachabwesenheit zu deklarieren. Dies ist unerwünscht, denn in diesen Zeit-Frequenz-Punkten wird das Sprachsignal unterdrückt, was zur unerwünschten Verzer-

rungen im Sprachsignal führt. Also schlägt [FRB07] zwei Verbesserungsvorschläge zum MCRA-Verfahren vor. Die erste Verbesserung betrifft die Realisierung der Minimumsuche mit überlappenden Fenstern, die dafür sorgt, dass die Rauschleistungsdichte mit kleinerer Verzögerung berechnet wird. Die zweite Verbesserung ist heuristischer Art und führt einen zusätzlichen Steuerungsparameter ein, der ähnlich wie beim IMCRA-Verfahren dafür sorgt, dass die Rauschleistungsdichte während der Zeit-Frequenz-Punkten mit relativ starker LDS nicht aktualisiert wird.

6. SNT: Das *subspace noise tracking* (SNT) Verfahren in [HJH08] basiert auf einer Eigenwertzerlegung der Kovarianzmatrix, die aus den komplexwertigen STFT-Koeffizienten des gestörten Signals berechnet wird. Diese Kovarianzmatrix wird aus den Kurzzeitspektren berechnet, die zunächst mittels einer Transformation weiß gemacht werden. Durch dieses Weißmachen werden allerdings nur die Korrelationen der Kurzzeitspektren eliminiert, die durch die STFT hervorgerufen werden. Die dafür notwendige Transformationsmatrix wird im Vorfeld zur Signalverarbeitung heuristisch ausgerechnet und zwar in Experimenten mit weißem Rauschen. Somit wird im Verfahren angenommen, dass die Korrelationen des Störsignals im Vergleich zu den Korrelationen, die durch die STFT bedingt sind, vernachlässigt werden können. Die resultierende Kovarianzmatrix wird mit Hilfe der Eigenwertzerlegung in zwei orthogonale Untermatrizen aufgeteilt, die jeweils die Räume mit und ohne Sprachsignal repräsentieren. Die Eigenwerte der letzten Untermatrix werden für die Schätzung der Rauschleistungsdichte verwendet und zwar auch in den Zeit-Frequenz-Punkten mit Sprachaktivität. Die Schätzung der Dimension der Untermatrix ohne Sprachsignal ist die Kernkomponente des SNT-Verfahrens. Da der resultierende Rauschschätzer biasbehaftet ist, wird eine Biaskompensation benötigt. Der dafür notwendige Biaskompensationsfaktor wird heuristisch bestimmt. Zusätzlich werden die RLDS-Schätzwerte mittels einer Rekursionsgleichung wie in (2.47) mit einem adaptiven Glättungsfaktor geglättet. Damit die RLDS-Schätzung auch in Anwesenheit von deterministischen Komponenten im Störsignal gut funktioniert, wird außerdem eine Minimumsuche ähnlich wie beim MS-Verfahren durchgeführt. Die große Stärke des SNT-Verfahrens ist seine Fähigkeit das LDS des Störsignals auch während der Sprachaktivität zu schätzen. Das SNT-Verfahren erreicht eine gute Leistungsfähigkeit in Gegenwart der nichtstationären Störung, benötigt allerdings dafür eine Trainigsphase und ist aufgrund der Eigenwertzerlegung, die in jedem Zeit-Frequenz-Punkt ausgeführt werden muss, sehr rechenintensiv.

7. MCRA-MAP: Das MCRA-Verfahren unterstützt vom *maximum a posteriori* (MAP) Schätzer aus [KPC09] greift einen Schwachpunkt des MCRA-Verfahrens hinsichtlich der SPP-Schätzung auf und versucht diesen zu verbessern. Beim MCRA-Verfahren ist es nämlich so, dass der Indikator der Sprachpräsenz nur auf Basis der aktuellen Beobachtung ausgerechnet wird. Da allerdings das Sprachsignal starke eigene Korrelationen zwischen den benachbarten Zeit-Frequenz-Punkten aufweist, sollte auch die Umgebung des aktuellen Zeit-Frequenz-Punktes in die SPP-Schätzung mit einbezogen werden. Und obwohl die erwähnten Korrelationen im Rahmen des MCRA-Verfahrens durch die rekursive Glättung der Indikatoren der Sprachpräsenz indirekt berücksichtigt werden, versucht [KPC09] durch die Verwendung einer MAP-basierten Schätzvorschrift eine robustere SPP-Schätzung zu erreichen. Dabei wird statt der traditionellen MAP-Entscheidungsregel das bedingte MAP-Kriterium verwendet, in das der SPP-Schätzwert des vorigen Zeit-Frequenz-Punktes als ein *a priori* Wissen einfließt. Im Unterschied zum MCRA-Verfahren wird außerdem eine zeitvariante Entscheidungsgrenze verwendet, die davon abhängig gemacht wird, ob im vorigen Zeit-

Frequenz-Punkt die Sprachpräsenz vermutet wurde oder nicht. Diese Wahl wird durch die Dünnbesetztheit des Sprachsignals im STFT-Bereich rechtfertigt, denn der Zustand der Abwesenheit des Sprachsignals wird von einem Zeit-Frequenz-Punkt zum nächsten nur ungern verlassen, im Vergleich zum Verlassen des Zustandes für die Sprachsignalanwesenheit. Diese Tatsache führte sowohl zu besseren Ergebnissen bei der SPP-Schätzung als auch zur besseren RLDS-Schätzung. Als Folge davon wird eine höhere Qualität der entstörten Sprachsignale insbesondere in Gegenwart der Störungen mit gemäßigter Instationarität registriert.

8. MMSE-VAD: Im Rahmen des MMSE-Verfahrens von Yu aus [Yu09] wird die Amplitude der STFT-Koeffizienten des Störsignals in einem Zeit-Frequenz-Punkt mit einer Rayleigh-Verteilung modelliert. Die *a posteriori* Verteilung dieser Amplitude, gegeben die STFT-Koeffizient des gestörten Signals im selben Zeit-Frequenz-Punkt, ist dann eine Rice-Verteilung [Ric48]. Daraus kann ein momentaner MMSE-Schätzwert der Rauschleistungsdichte berechnet werden, der sowohl vom *a priori* als auch vom *a posteriori* SNR abhängt, die zunächst unbekannt sind und im Rahmen des Verfahrens geschätzt werden müssen. Während das *a priori* SNR mit Hilfe des DD-Verfahrens aus [EM84] berechnet wird, wird die Rauschleistungsdichte über eine rekursive Glättung der momentanen MMSE-Schätzwerte mit einem konstanten Glättungsparameter von $0{,}96$ berechnet, die einer Verfolgungsgeschwindigkeit von etwa $v_P = 22.2\,\text{dB/s}$ entspricht. Um eine Überschätzung der Rauschleistungsdichte zu vermeiden, die der fehlerhaften Schätzung des *a priori* SNR geschuldet ist, die bei den sogenannten *onsets* des Sprachsignals verstärkt auftritt, wird der MMSE-Schätzwert samt der rekursiven Glättung nur in den Zeit-Frequenz-Punkten ausgerechnet, in denen das LDS des gestörten Signals nicht all zu sehr von der zuletzt geschätzten Rauschleistungsdichte unterscheidet. Wird im aktuellen Zeit-Frequenz-Punkt einen plötzlichen Anstieg der Signalleistung beobachtet, der eine heuristisch definierte Schranke übersteigt, wird die geschätzte Rauschleistungsdichte konstant gehalten. Aufgrund der zuletzt erwähnten Modifikationen verliert der MMSE-Schätzer seine Erwartungstreue und bedarf einer Biaskompensation, die vor der ausgangsseitigen Glättung durchgeführt werden kann. Im Rahmen dieser Arbeit allerdings verzichten wir auf die Biaskompensation, denn sie wird in [Yu09] als nicht obligatorisch angesehen. Das resultierende Verfahren ist recheneffizient, leistungsfähig und einfach zu parametrisieren, beinhaltet jedoch Heuristiken, die nicht für alle Szenarien generalisierbar sein können.

9. MMSE-BM[3]: Im MMSE-basierten Rauschschätzer von Hendriks aus [HHJ10] wird dieselbe statistische Modellierung wie in [Yu09] verfolgt, was auch im gleichen MMSE-Schätzer der momentanen Rauschleistungsdichte resultiert. Allerdings wird in [HHJ10] im Unterschied zum [Yu09] eine geschlossene analytische Lösung für die Biaskompensation vorgestellt, die dadurch ermöglicht wird, dass der Biaskompensationsfaktor für den ML-Schätzwert des *a priori* SNR hergeleitet wird. In der Realisierung des Verfahrens kommt aber auch das DD-Verfahren zum Einsatz. Um die Varianz der MMSE-Schätzwerte der Rauschleistungsdichte zu reduzieren, werden die Schätzwerte nach der Biaskorrektur rekursiv mit einem konstanten Glättungsparameter von $0{,}8$ geglättet, die mit einer relativ hohen Verfolgungsgeschwindigkeit von etwa $v_P = 60.6\,\text{dB/s}$ einhergeht. Da die Glättungskonstante einen relativ kleinen Wert hat, ist der MMSE-Schätzer von Hendriks prinzipiell im Stande, der zeitvarianten Rauschleistungsdichte zu folgen. Nur in dem Fall eines abrupt steigenden RLDS kann es zum Erliegen der Verfolgungseigenschaft des Schätzers kommen. Um diesen negati-

[3]Die Abkürzung BM steht hier für die Techniken der Biaskompensation und der Minimumsuche, die in diesem Verfahren zum Einsatz kommen.

ven Effekt zu vermeiden, wird die Methode des sogenannten Sicherheitsnetzes (engl. *safety net*) angewandt, die in [EH08] eingeführt wird. Dabei wird eine aktuelle untere Schranke für die Rauschleistungsdichte berechnet, unter die der MMSE-Schätzwert nicht fallen darf. Diese Schranke wird im Rahmen einer Minimumsuche über die vergangenen Werte des gestörten Spektrogramms berechnet, die allerdings nicht weiter als $0{,}8$ Sekunden in der Vergangenheit liegen. Somit wird der zurückgefallene MMSE-Schätzwert aufgefangen und in den Bereich der tatsächlichen Werte der Rauschleistungsdichte gebracht, falls es zum Erliegen der Verfolgungseigenschaft des Schätzers kommen soll. Als Folge entsteht in [HHJ10] ein robuster Rauschschätzer, der in den experimentellen Untersuchungen in [TTM+11] im Vergleich zu einigen anderen Rauschschätzern gute Leistungsfähigkeit zeigt.

10. SPP-FP: In [GH11] wird der MMSE-BM-Schätzer aus [HHJ10] analysiert und neu interpretiert. Daraus entsteht ein RLDS-Schätzer, der fünf Verarbeitungsschritte beinhaltet und eine ähnliche Strategie wie der VAD-RA-Schätzer aus [Hir93] verfolgt, allerdings mit dem Unterschied, dass statt einer harten VAD-Entscheidung eine weiche wertekontinuierliche SPP-Schätzung zum Einsatz kommt, die in der Regel Schätzwerte im Bereich $[0; 1]$ liefert. Diese wird im ersten Schritt mit Hilfe einer fixierten *a priori* SNR (engl. *fixed priors*, FP) berechnet, die bei der Entwicklung des Schätzers festgelegt (und nicht mehr geschätzt) wird [GBM08]. Wie in [GH11] argumentiert, entsteht dadurch ein erwartungstreuer RLDS-Schätzer, der weder eine Biaskompensation noch eine Verwendung der *safety net* Methode bedarf. Allerdings ist dieser SPP-FP-Schätzer bei kleinen Werten von $|Y(k,\ell)|^2$ und Unterschätzung von $\hat{\lambda}_D(k,\ell)$ gefährdet, in einen Stillstand zu geraten[4]. Um den Stillstand zu vermeiden, wird die berechnete SPP-Schätzung (ähnlich wie beim MCRA-Schätzer aus [CB01a]) im zweiten Schritt rekursiv mit der Glättungskonstante 0.9 geglättet, was einer Verfolgungsgeschwindigkeit von etwa $v_F = 57.2\,\text{dB/s}$ entspricht. Im dritten Schritt bekommt der SPP-Schätzwert eine Obergrenze von 0.99, falls die geglättete SPP-Schätzung aus dem zweiten Schritt denselben Wert überschreitet. Die korrigierten SPP-Schätzwerte werden im vierten Schritt für die Berechnung der ungeglätteten RLDS-Schätzwerte verwendet. Um die Schätzfehlervarianz zu reduzieren, die durch eine fehlerbehaftete SPP-Schätzung entstehen kann, werden die aus der SPP-Schätzung resultierenden RLDS-Schätzwerte im fünften Schritt rekursiv mit einer Glättungskonstante 0.8 (geht mit der Verfolgungsgeschwindigkeit von etwa $60.6\,\text{dB/s}$ einher) geglättet, bevor sie an den Ausgang des Schätzers weitergeleitet werden. Laut den experimentellen Untersuchungen in [GKR12] ist der SPP-FP-Schätzer im Vergleich zum MMSE-BM-Schätzer zum einen weniger anfällig im Bezug auf die Unterschätzung von RLDS und zum anderen recheneffizienter.

An dieser Stelle soll erwähnt werden, dass nur vier der vorgestellten Rauschschätzer in der Lage sind, das LDS des Störsignals permanent (auch während der Sprachaktivität) zu schätzen. Das sind das OSMS-Verfahren, das SNT-Verfahren und die beiden MMSE-Schätzer. Die anderen Rauschschätzer aktualisieren ihre Schätzwerte während der Sprachaktivität nicht und halten sie auf dem zuletzt geschätzten Niveau.

3.1.2. Auflistung verwendeter Techniken

Wie die vorhergehenden Ausführungen zeigten, bringen die betrachteten RLDS-Schätzer unterschiedliche Komponentenkomplexität mit sich. So benötigen 7 der 10 betrachteten RLDS-

[4] Dabei liefert die SPP-Schätzung fälschlicherweise eine 1 und signalisiert damit Sprachsignalpräsenz, obwohl das Sprachsignal eigentlich abwesend ist und die SPP-Schätzung eine 0 ausgeben soll [GH11].

Schätzer eine eigenständige SPP-Schätzung, die nicht mit dem Baustein 4 des Systems in Abb. 2.2 verwechselt werden darf. Außerdem machen die beiden MMSE-basierten Schätzer sogar von einer internen *a priori* SNR-Schätzung Gebrauch, die sich vom Baustein 3 des Systems in Abb. 2.2 unterscheidet und intern in dem jeweiligen RLDS-Schätzer realisiert wird. Während der MMSE-BM-Schätzer eine ML-Schätzung der *a priori* SNR realisiert, benutzt der MMSE-VAD-Schätzer das DD-Verfahren dafür. Außerdem sind die verwendeten MMSE-Schätzer hier nichts anderes als die spektralen Filterfunktionen, die für die momentane spektrale Rauschleistung $|D(k,\ell)|^2$ gegeben die verrauschte Beobachtung $Y(k,\ell)$ hergeleitet werden. Somit tendieren die beiden MMSE-Schätzer der spektralen Rauschleistungsdichte dazu, als autonome Untersysteme zu agieren. Im Gegensatz dazu benötigen die MS-basierten Rauschschätzer aus [Mar94] und [Mar01] keine derartigen Komponenten. Ihre Unabhängigkeit von einer SPP-Schätzung trägt im besonderen Maße zur ihrer großen Popularität bei. Allerdings lässt die Leistungsfähigkeit der MS-basierten Verfahren in der Gegenwart der nichtstationären Störungen etwas nach.

Beim Studieren der modernen RLDS-Schätzer fällt auf, dass manche Techniken in den unterschiedlichen Schätzern immer wieder zum Einsatz kommen. Beim genaueren Durcharbeiten kristallisieren sich fünf Grundtechniken heraus, die auf bestimmten Eigenschaften der gestörten Sprachsignale aufbauen und in Tab. 3.1 zusammengefasst sind. Hier ist auch angegeben, welche Techniken bei welchen der vorgestellten Schätzer zum Einsatz kommen. Im folgenden werden diese Techniken der Reihe nach erläutert.

1. VAD/SPP Schätzung: 7 der 10 betrachteten RLDS-Schätzer greifen entweder auf eine VAD-Schätzung oder auf eine SPP-Schätzung zurück, um die Zeit-Frequenz-Punkte mit und ohne Sprachaktivität zu finden [HE95, CB01a, Coh03, FRB07, KPC09, Yu09, GKR12]. Anhand dieser Information aktualisieren sie ihre Schätzwerte nur in den Zeit-Frequenz-Punkten ohne Sprachaktivität und sonst halten sie sie meistens auf einem konstanten Wert.

	Name	Quelle	VAD/SPP Schätzung	Minimumsuche	Biaskompensation	Bayessche Inferenz	Ausgangsglättung
1.	VAD-RA	[Hir93, HE95]	✓				✓
2.	OSMS	[Mar01]		✓	✓		
3.	MCRA	[CB01a, CB02]	✓	✓			✓
4.	IMCRA	[Coh03]	✓	✓	✓		✓
5.	EMCRA	[FRB07]	✓	✓			✓
6.	SNT	[HJH08]		✓	✓		✓
7.	MCRA-MAP	[KPC09]	✓	✓			✓
8.	MMSE-VAD	[Yu09]	✓			✓	✓
9.	MMSE-BM	[HHJ10]		✓	✓	✓	✓
10.	SPP-FP	[GH11, GKR12]	✓				✓

Tabelle 3.1.: Wichtige Techniken moderner RLDS-Schätzer.

2. Minimumsuche: Aufgrund der Dünnbesetztheit der ungestörten Sprachsignale im ZF-Bereich sinkt die momentane spektrale Leistung des verrauschten Signals $|Y(k,\ell)|^2$ regelmäßig auf das Niveau der momentanen spektralen Leistung des Störsignals $|D(k,\ell)|^2$, das in direkter Verbindung mit dem gesuchten RLDS-Wert steht. Aus diesem Grund führen 7 der 10 betrachteten RLDS-Schätzer eine kausale Minimumsuche im LDS-Bereich über eine bestimmte Anzahl der vergangenen Rahmen, die mit dem Signaldauer von etwa einer Sekunde einhergeht [Mar01, CB01a, Coh03, FRB07, HJH08, KPC09, HHJ10].

3. Biaskompensation: Da das gestörte Sprachsignal häufig als ein Zufallsprozess modelliert wird, lassen sich resultierende Schätzer auf die Erwartungstreue überprüfen. Ist ein Schätzer nicht erwartungstreu, gelingt es häufig, den Bias des Schätzers zu bestimmen, um ihn im Rahmen einer Biaskompensation zu eliminieren. Demzufolge führen 4 der 10 betrachteten RLDS-Schätzer eine Biaskompensation durch [Mar01, Coh03, HHJ10].

4. Bayessche Inferenz: Außerdem erlaubt eine statistische Modellierung die Durchführung einer statistischen Inferenz, wie dies bei den beiden MMSE-basierten RLDS-Schätzern der Fall ist [Yu09, HHJ10]. Die Verwendung der MMSE-Schätzung kann man auch als die Anwendung einer spektralen Filterfunktion auf das Spektrogramm des gestörten Signals $|Y(k,\ell)|^2$ ansehen, die zum Ziel hat, daraus das ungestörte Signal zu entfernen.

5. Ausgangsglättung: Als letzte Technik, die jedoch in der RLDS-Schätzung am weitesten verbreitet ist, stellt sich die Ausgangsglättung dar. Definitionsgemäß wird sie als letzter Verarbeitungsschritt in 9 der 10 vorgestellten RLDS-Schätzer realisiert [HE95, CB01a, Coh03, FRB07, HJH08, KPC09, Yu09, HHJ10, GKR12]. Meistens wird dafür die Rekursionsgleichung erster Ordnung wie in (2.48) verwendet. Der Grund für große Popularität der Ausgangsglättung ist die Tatsache, dass das weiße Rauschen als eine einzige Störung in der Realität selten vorkommt und dass die meisten akustischen Störsignale starke Korrelationen aufweisen. Ein weiterer Argument für die Verwendung der ausgangsseitigen Glättung ist die Signalverarbeitung mit überlappenden Rahmen, bei dem die benachbarten STFT-Koeffizienten selbst beim weißen Rauschen korreliert sind[5].

Zu rekursiven Glättung lässt sich zusätzlich hinzufügen, dass diese in der RLDS-Schätzung nicht nur für die Ausgangsglättung verwendet wird, sondern auch in zwei weiteren Techniken, sodass insgesamt drei verschiedene Arten der Glättung unterschieden werden können. Neben der Ausgangsglättung gibt es noch eine Glättung im Rahmen der SPP-Schätzung und bei der Minimumsuche. Bei der SPP-Schätzung wird die rekursive Glättung mit einem konstanten Glättungsparameter eingesetzt, um die Schätzfehlervarianz der SPP-Schätzung zu reduzieren. Aufgrund der hohen Instationarität von Sprachsignalen wird bei der Minimumsuche ein zeitvarianter Glättungsparameter benötigt.

3.2. Generalisierte spektrale Filterfunktionen

Einige konventionelle in der spektralen Sprachsignalentstörung weit verbreitete Filterfunktionen wie das Wiener-Filter, der MMSE-LSA-Schätzer und die OMLSA-Filterfunktion wurden in Abschnitt 2.2 bereits vorgestellt. Da im Fokus dieser Arbeit die Entwicklung von Verfahren zur generalisierten modellbasierten Sprachsignalentstörung steht, wird in diesem Abschnitt die generalisierte spektrale Sprachsignalentstörung betrachtet, wie die generalisierte

[5] Allerdings können diese Korrelationen vernachlässigt werden, wenn die Überlappung der aufeinander folgenden Blöcke weniger als eine halbe Fensterlänge ist [Coh05].

spektrale Subtraktion (engl. *generalized spectral subtraction*, GSS) und die generalisierten modellbasierten Verfahren. Das Interesse an der GSS wird zuletzt durch die Untersuchungen in [ITS$^+$10, IST$^+$11] neu geweckt. Hier wird gezeigt, dass es eigentlich keine theoretische Rechtfertigung für die Entstörung der Sprachsignale im Bereich der spektralen Amplituden $|Y(k,\ell)|$ und Leistungen $|Y(k,\ell)|^2$ gibt. Trotzdem beschäftigt man sich in etwa $90\,\%$ der Veröffentlichungen mit der Signalentstörung in diesen Bereichen. Die objektiven und subjektiven Untersuchungen in [IST$^+$11] verdeutlichen allerdings, dass die Entstörung der spektralen Wurzelamplituden vorteilhafter ist hinsichtlich einer hoch-qualitativen Rauschunterdrückung mit wenig *musical noise*. Außerdem wird die Potenzierung von spektralen Amplituden mit einem beliebigen Exponenten in der automatischen Spracherkennung erfolgreich eingesetzt [RAS06, DP17].

Der Ausgangspunkt der konventionellen generalisierten spektralen Subtraktion ist die Annahme der Additivität der generalisierten spektralen Amplituden:

$$|Y(k,\ell)|^\beta = |S(k,\ell)|^\beta + |D(k,\ell)|^\beta \tag{3.1}$$

mit einem Kompressionsfaktor $\beta \in \mathbb{R}_{>0}$. Während die Additivität der STFT in der Gleichung (2.3) aufgrund der Linearität der STFT exakt gilt, ist die Gleichung (3.1) nur eine Approximation. Die Güte dieser Approximation für die Werte $\beta \in [0.5;\ 2]$ wird in [Vor15] hinsichtlich der Qualität der entstörten Signale gemessen in MOS-LQO$_{\text{WB}}$ untersucht. Dabei wird die generalisierte spektrale Amplitude des entstörten Signals mit der einfachen Subtraktionsregel berechnet:

$$|\hat{S}(k,\ell)|^\beta = \max\left(\,|Y(k,\ell)|^\beta - |D(k,\ell)|^\beta\,,\,0\,\right), \tag{3.2}$$

wobei die wahre spektrale Amplitude des Störsignals $|D(k,\ell)|$ als bekannt vorausgesetzt wird. Die Studie kommt zu einigen interessanten Ergebnissen. Bei einer Entstörung von Sprachsignalen in Gegenwart von nichtstationären Störungen wird im Mittel die beste Signalqualität für $\beta \approx 1$ erreicht, wobei der optimale Kompressionsfaktor mit dem steigenden globalen eingangsseitigen SNR-Wert leicht wächst und zwar von 0.95 für $\text{SNR}_{\text{IN}} = -10\,\text{dB}$ auf 1.15 für $\text{SNR}_{\text{IN}} = 40\,\text{dB}$. Beinhaltet die Störung musikalische Instrumente, soll laut [Vor15] der Kompressionsfaktor $\beta > 1.05$ gewählt werden. Zum Beispiel sind für 4, 3 oder 2 Instrumente jeweils die Werte 1.1, 1.25 und 1.35 optimal. Aber nicht nur Werte $\beta > 1$ sind für die Signalentstörung vorteilhaft, wie Experimente mit dem tiefpasslastigen Rauschen zeigten, das auf der Straße aufgenommen wurde. Hier ist es vorteilhaft einen frequenzabhängigen Kompressionsfaktor β zu verwenden, der die Werte von 0.2 bei tiefen Frequenzen bis 1.25 bei hohen Frequenzen annimmt.

Die spektralen Amplituden $|D(k,\ell)|$ sind in der Realität jedoch unbekannt und werden durch den Schätzwert des generalisierten RLDS $E[|D(k,\ell)|^\beta]$ ersetzt, das wie in [STCT98] gezeigt aus dem RLDS $\lambda_D(k,\ell)$ folgendermaßen berechnet werden kann:

$$E[|D(k,\ell)|^\beta] = \Gamma\left(\frac{\beta}{2}+1\right)\cdot\lambda_D^{\frac{\beta}{2}}(k,\ell), \tag{3.3}$$

wobei $\Gamma(x)$ die Gamma-Funktion ist. Dadurch entsteht die generalisierte spektrale Subtraktion, die bereits in [BSM79] folgendermaßen eingeführt wird:

$$|\hat{S}(k,\ell)|^\beta = \max\left(\,a\cdot|Y(k,\ell)|^\beta - b\cdot\hat{\lambda}_D^{\frac{\beta}{2}}(k,\ell)\ ,\ c\cdot\hat{\lambda}_D^{\frac{\beta}{2}}(k,\ell)\,\right), \tag{3.4}$$

wobei a ein Normalisierungsfaktor (engl. *normalization factor*), b ein Übersubtraktionsfaktor (engl. *over-subtraction factor*) und c ein spektraler Bodenparameter (engl. *spectral floor parameter*) sind. Bei den Untersuchungen mit drei verschiedenen Werten von $\beta = \{0.5,\, 1,\, 2\}$ stellte man die beste Signalqualität bei $\beta = 2$ fest, also wenn die GSS in die spektrale Leistungssubtraktion wie in [CN78] übergeht, wo sie als *correlation subtraction method* bezeichnet wird. Man beachte, dass bei $\beta = 1$ die GSS (3.4) der spektralen Subtraktion aus [Bol79] entsprechen würde, die in Kap. 1 bereits erwähnt wurde[6]. Eine der ersten Publikationen, in der die generalisierten spektralen Amplituden statistisch beschrieben werden, ist [WAP75]. Hier werden die Kompressionsfaktoren $\beta \in \{\frac{1}{4},\, \frac{1}{3},\, \frac{1}{2},\, 1\}$ vorgeschlagen, sodass die Signalverarbeitung auf den spektralen Wurzelamplituden stattfindet. In [Lim78] und [LO79] wird beobachtet, dass die mit (3.4) entstörten Signale für $\beta \in \{0.5,\, 1\}$ stärkere Rauschunterdrückung aufweisen. Wenn GSS für eine ASR-Aufgabe eingesetzt werden soll, wird in [KGW$^+$89] der Kompressionsfaktor zu $\beta = 0.5$ gewählt.

Alternativ lässt sich der GSS-Schätzer aus (3.4) mit Hilfe einer spektralen Filterfunktion $G_\beta^{\text{GSS}}(k,\ell)$ ausdrücken in der Form

$$\hat{S}_\beta(k,\ell) = G_\beta^{\text{GSS}}(k,\ell) \cdot Y_\beta(k,\ell), \tag{3.5}$$

wobei aus Gründen der besseren Übersichtlichkeit eine Abkürzung eingeführt wird

$$X_\beta(k,\ell) = |X(k,\ell)|^\beta \;\text{ für }\; X \in \{Y, S, D\}. \tag{3.6}$$

Generalisierte spektrale Amplituden (GSA) $Y_\beta(k,\ell)$ werden also mit einer GSS-Filterfunktion $G_\beta^{\text{GSS}}(k,\ell)$ entstört, die mit (3.4) für $a = 1$ und $b = \Gamma\left(\frac{\beta}{2}+1\right)$ folgende Form hat:

$$G_\beta^{\text{GSS}}(k,\ell) = 1 - \frac{\Gamma\left(\frac{\beta}{2}+1\right)}{\gamma^{\frac{\beta}{2}}(k,\ell)}. \tag{3.7}$$

Modellbasierte Filterfunktionen: Lange Zeit legte man die Parameter a, b und c aus (3.4) heuristisch entsprechend den Ergebnissen experimenteller Untersuchungen fest. Jedoch lassen sie sich mit Hilfe einer statistischen Modellierung eleganter bestimmen. So werden die Parameter a und b in [STCT98] für einen parametrischen GSS (PGSS) Schätzer in der Form $|\hat{S}(k,\ell)|^\beta = a \cdot |Y(k,\ell)|^\beta - b \cdot E[|D(k,\ell)|^\beta]$ im MMSE-Sinne über die Minimierung der Kostenfunktion $E[(|S(k,\ell)|^\beta - |\hat{S}(k,\ell)|^\beta)^2]$ gewählt. Unter der Annahme, dass $S(k,\ell)$ und $D(k,\ell)$ in (2.3) unkorreliert sind und jeweils den Verteilungsdichtefunktionen (2.4) und (2.5) folgen, resultiert eine generalisierten modellbasierte Entstörungsregel:

$$\hat{S}_\beta(k,\ell) = E\left[\, S_\beta(k,\ell) \,|\, Y_\beta(k,\ell) \,\right] = G_\beta^{\text{PGSS}}(k,\ell) \cdot Y_\beta(k,\ell) \tag{3.8}$$

mit einer spektralen modellbasierten PGSS-Filterfunktion im GSA-Bereich

$$G_\beta^{\text{PGSS}}(k,\ell) = \frac{\xi^\beta(k,\ell)}{\xi^\beta(k,\ell)+1} \cdot \left(1 - \frac{\Gamma\left(\frac{\beta}{2}+1\right)}{\gamma^{\frac{\beta}{2}}(k,\ell)} \left(1 - \xi^{-\frac{\beta}{2}}(k,\ell)\right)\right). \tag{3.9}$$

[6] In [PWS10] wird auch die spektrale Subtraktion nach [BSM79] für $\beta = \{1,\, 2\}$ im sogenannten Kurzzeit-Modulation (engl. *short-time modulation*, STM) Bereich realisiert. Dabei wird der zeitliche Verlauf von $|Y(k,\ell)|$ für jedes Frequenzband mit einer zusätzlichen STFT in den STM-Bereich überführt und dort mit der spektralen Subtraktion entstört. Die Autoren berichten von einer signifikanten Verbesserung der Qualität der entstörten Signale bei Verwendung der spektralen Leistungssubtraktion für $\beta = 2$.

Während der Term vor der großen Klammer die Form des generalisierten Wiener-Filters aus [LO79] hat, ähnelt der zweite der GSS-Filterfunktion aus (3.7). Prägt man dem PGSS-Schätzer eine Randbedingung $b = a$ ein, resultiert die MMSE-Optimierung in

$$G_\beta^{\text{PGSSR}}(k,\ell) = \frac{\xi^\beta(k,\ell)}{\xi^\beta(k,\ell) + c_\beta} \cdot \left(1 - \frac{\Gamma\left(\frac{\beta}{2}+1\right)}{\gamma^{\frac{\beta}{2}}(k,\ell)}\right), \tag{3.10}$$

wobei $c_\beta = 1 - \Gamma^2(\frac{\beta}{2}+1)/\Gamma(\beta+1)$ eine von β abhängige Konstante ist. Der zweite Term im Produkt (3.10) wird dann exakt zur Filterfunktion $G_\beta^{\text{GSS}}(k,\ell)$ aus (3.7). Laut Untersuchungen in [STCT98] für $\beta = \{1,\, 2\}$ zeigen die Filterfunktionen PGSS und PGSSR im Vergleich zum GSS-Schätzer eine verbesserte Störsignaldämpfung. Die PGSSR-Filterfunktion ist dabei konkurrenzfähig mit dem MMSE-LSA-Schätzer aus (2.27). Die Eigenschaften der PGSS-Filterfunktion lassen sich verbessern, falls ein adaptiver Kompressionsfaktor im Wertebereich $[0.5;\, 2]$ verwendet wird, der vom lokalen SNR gesteuert wird [LSH$^+$08].

Eine weitere generalisierte modellbasierte spektrale Funktion wird in [YKR03, YKR05] auch durch die Berechnung des Erwartungswertes $E\,[\,S_\beta(k,\ell)\,|\,Y_\beta(k,\ell)\,]$ hergeleitet, allerdings ohne eine parametrische Anforderung an die Form des Schätzers, wie dies in [STCT98] der Fall ist. Daraus resultiert folgende MMSE-GSS-Filterfunktion:

$$G_\beta^{\text{MMSE-GSS}}(k,\ell) = \Gamma\left(\frac{\beta}{2}+1\right) \cdot M\left(-\frac{\beta}{2};\, 1;\, -v(k,\ell)\right) \cdot \frac{v^{\frac{\beta}{2}}(k,\ell)}{\gamma^\beta(k,\ell)}, \tag{3.11}$$

wobei $M(x;y;z)$ die konfluente hypergeometrische Funktion ist. Der MMSE-GSS-Schätzer hat eine bemerkenswerte Eigenschaft: während er für $\beta = 1$ in den MMSE-SA-Schätzer aus [EM84] übergeht, konvergiert er für $\beta \to 0$ gegen den MMSE-LSA-Schätzer aus [EM85]. Um diese Eigenschaft für eine adaptive Filterung gewinnbringend ausnutzen zu können, wird in [YKR05] vorgeschlagen, den Kompressionsfaktor β im Wertebereich $\beta \in (0;\, 4]$ ähnlich wie in [LSH$^+$08] mit Hilfe eines segmentellen SNR zu adaptieren. In experimentellen Untersuchungen in [LSH$^+$08] übertrifft die adaptive MMSE-GSS-Filterfunktion den MMSE-LSA-Schätzer aus (2.27) sowohl hinsichtlich der Qualität der prozessierten Signale als auch bezüglich Störsignaldämpfung. Allerdings ist der große Nachteil der MMSE-GSS-Filterfunktion ihre hohe Rechenkomplexität aufgrund ihrer Abhängigkeit von der Funktion $M(x;y;z)$, die allerdings für eine recheneffiziente Realisierung tabellarisiert werden kann. Um die Leistungsfähigkeit der MMSE-GSS-Funktion zu gewährleisten, soll außerdem laut [LSH$^+$08] das RLDS systematisch überschätzt werden, denn sonst funktioniert die vorgeschlagene Adaption des Kompressionsfaktors nicht wie gewünscht. Eine nennenswerte Erweiterung der MMSE-GSS-Filterfunktion wird in [PC07b] vorgeschlagen, wo neben positiven auch negative Werte von β zugelassen werden. In den experimentellen Untersuchungen stellt sich hier heraus, dass der MMSE-GSS-Schätzer aus (3.11) für $\beta = -1$ leicht bessere Qualität als der MMSE-LSA-Filterfunktion aus (2.27) liefert.

Verwendung von generalisierten Verteilungsdichtefunktionen: Generalisierung spektraler Amplituden ist allerdings nicht der einzige Weg, um zu den generalisierten Filterfunktionen zu gelangen. Ein anderer Weg ist die Modellierung von spektralen Größen mit generalisierten Verteilungen, die sich über ihre Parameter (wie ein Formparameter oder ein Freiheitsgrad) an die experimentelle Verteilung der Daten anpassen. So können die spektralen Amplituden $|S(k,\ell)|$ mit Verteilungen modelliert werden, die sie besser als eine Rayleigh-Verteilung beschreiben. In [DTI05] führt die Verwendung einer generalisierten Gamma-Verteilung zu einem MAP-basierten Schätzer, der in den experimentellen Untersuchungen

im Vergleich zu dem Wiener-Filter (2.24) und zur LSA-Filterfunkton (2.27) bessere Leistungsfähigkeit zeigt. Allerdings gelingt es hier nicht, eine geschlossene Form für die generalisierte Filterfunktion zu finden. Aus diesem Grund wird sie nummerisch mit Hilfe des Newton-Raphson Verfahrens berechnet. Auch der Real- und Imaginärteile von $S(k,\ell)$ können mit steilgipfligen Verteilungen modelliert werden wie in [Mar05]. Hier werden dafür zum Einen eine komplexe Laplace-Verteilung zum Anderen eine komplexe bilaterale Gamma-Verteilung verwendet, welche steilgipflig (engl. *super gaussian*) sind. Im Vergleich zu den Schätzern aus [EM84] und [EM85] erreichten die resultierenden MMSE-basierten Filterfunktionen größere Werte des segmentellen SNR und kleinere Werte der spektralen Verzerrung. Allerdings blieben die Schätzer [EM84] und [EM85] hinsichtlich der Qualität des restlichen Rauschens ungeschlagen. In [LV05] werden die spektralen Amplituden mit einer generalisierten steilgipfligen Verteilung modelliert, woraus zwei MAP-basierte Filterfunktionen resultieren. Wie experimentelle Untersuchungen in [LV05] zeigten, führte die an die Daten angepasste Modellierung zu prozessierten Signalen besserer Qualität im Vergleich zur MMSE-SA-Filterfunktion aus [EM84].

In [AW06] werden $|S(k,\ell)|$ entweder mit einer Gamma- oder mit einer Chi-Verteilung modelliert und sowohl zwei generalisierte MMSE- als auch zwei generalisierte MAP-basierte Schätzer hergeleitet. Bei Gegenüberstellung von diesen Verfahren mit den MMSE-SA- und MMSE-LSA-Schätzern in [AW09] zeigten die prozessierten Sprachsignale der generalisierten Schätzer ähnlicher Signalqualität bei etwas besserer Störsignaldämpfung. Mit einer ähnlichen Modellierung wie in [DTI05] werden in [EHHJ07] generalisierte MMSE-Schätzer entwickelt[7], die im Vergleich zum MMSE-SA-Schätzer aus [EM84] zu einer Verbesserung der Signalqualität führten[8]. Allerdings werden in [EHHJ07] im Unterschied zu [DTI05] der Realteil von $S(k,\ell)$ mit einer zweiseitigen generalisierten Gamma-Verteilung modelliert und daraus resultierende MMSE-Schätzer entwickelt. [WSST12] greift auch auf die statistische Modellierung aus [AW06] zurück und entwickelt eine Methode für die Adaption des Formparameters der verwendeten Verteilung. Und da der Formparameter über die Kurtosis der spektralen Amplituden $|S(k,\ell)|$ berechnet werden kann, wird die Kurtosis mit Hilfe der *moment-cumulant* Transformation berechnet. Die daraus resultierende generalisierte adaptive MMSE-Filterfunktion übertraf die Filterfunkton aus [AW06] ohne Adaption in experimentellen Untersuchungen mit Testpersonen hinsichtlich der Signalqualität.

Die Verwendung von generalisierten spektralen Amplituden bei gleichzeitiger Modellierung mit generalisierten Verteilungen wird zum ersten Mal in [BKM08] eingesetzt. Dabei wird hier ein generalisierter MMSE-Schätzer hergeleitet, bei dem sowohl der Kompressionsfaktor $\beta \in [-0.5;\ 1.5]$ als auch der Formparameter der Chi-Verteilung aus [AW06], der im Wertebereich $[0.1;\ 1]$ variiert wird, unabhängig voneinander gewählt werden. Die Optimierung der beiden Parameter führte zum Ergebnis, dass der Formparameter gleich dem Kompressionsfaktor $\beta = 0.5$ gewählt werden soll, woraus eine *Super Gaussian Amplitude Root* (SuGAR) Filterfunktion resultierte. In Experimenten erreichte der generalisierte SuGAR-Schätzer um durchschnittlich $0.5\,\mathrm{dB}$ bessere Werte des segmentellen SNR im Vergleich zu der Wiener-, der MMSE-SA- und der MMSE-LSA-Filterfunktionen, die in etwa ähnliche Leistungsfähigkeit zeigten. In [BM10] wird die SuGAR-Filterfunktion für

[7] In [EHHJ07] wird für die Modellierung von spektralen Amplituden genauso wie in [DTI05] die generalisierte Gamma-Verteilung verwendet, die allerdings anders parametrisiert wird.

[8] Leider nimmt der MMSE-LSA-Schätzer aus [EM85], der zur besseren Signalqualität als der MMSE-SA-Schätzer aus [EM84] führt, in experimentellen Untersuchungen weder in [LV05] noch in [EHHJ07] teil.

eine *a priori* SNR-Schätzung eingesetzt und die statistische Modellierung mit der dazugehörigen Chi-Verteilung für eine SPP-Schätzung verwendet. Außerdem ist hier ein guter Überblick über die generalisierten Filterfunktionen zu finden. Außerdem werden MMSE-Schätzer der log-spektralen Amplituden entwickelt, die auf generalisierten Modellen beruhen [BA11, ZZZLy12]. So werden in [BA11] die spektralen Amplituden mit der mächtigen generalisierten Gamma-Verteilung modelliert und zwei approximative LSA-Schätzer hergeleitet. In [ZZZLy12] wird einer der LSA-Schätzer aus [BA11] modifiziert und mit einer SPP-Schätzung multipliziert, die aus der statistischen Modellierung mit der generalisierten Gamma-Verteilung folgt. In den Experimenten mit den Filterfunktionen aus [EM84, CB01b, BA11] zeigt die vorgeschlagene Filterfunktion bessere Leistungsfähigkeit.

Generalisierte MAP-Schätzung: Eine besonderer Schätzer wird in [STWJ13] vorgestellt. Hier wird die Berechnungsvorschrift des MAP-Schätzers $|\hat{S}(k,\ell)| = \operatorname{argmax}_s \; p(y|s) \cdot [p(s)]^\beta$ generalisiert, indem die *a priori* Verteilungsdichtefunktion der ungestörten spektralen Amplituden $p(s)$ in der Schätzvorschrift mit einem Kompressionsfaktor β potenziert wird. Über β kann man den relativen Einfluss von $p(s)$ gegenüber $p(y|s)$ steuern. Dabei wird ein für eine Äußerung konstanter Kompressionsfaktor β vorgeschlagen, der entsprechend dem globalen eingangsseitigen SNR_{IN} mit Hilfe einer Sigmoid-Funktion gesetzt wird. Die resultierende Filterfunktion übertrifft bei Evaluierung den MMSE-SA-Schätzer aus [EM84] und zwei weitere MAP-basierte Filterfunktionen. In [TL16] wird die Filterfunktion aus [STWJ13] hinsichtlich der Wahl des Parameters β weiterentwickelt, der hier adaptiv gewählt wird. Die Optimierung von β bezüglich eines *Speech Distortion Index* Maßes aus [CBHD06] zeigte, dass der optimale Kompressionsfaktor β mit steigendem SNR_{IN} von $\beta = 2.0$ für $\text{SNR}_{\text{IN}} = 0\,\text{dB}$ auf $\beta = 0.5$ für $\text{SNR}_{\text{IN}} = 20\,\text{dB}$ sinkt.

Die Eigenschaften der beschriebenen generalisierten Filterfunktionen sind in Tab. 3.2 zusammengefasst. Während nur drei der generalisierten Schätzer von generalisierten spektralen

	Quelle	Gen. Amplitude	Gen. Verteilung	Gen. Vorschrift	MMSE	MAP	Rayleigh	Gamma	Chi	Gen. Gamma	Log-Schätzer
1.	[STCT98]	✓			✓		✓				
2.	[YKR03]	✓			✓		✓				
3.	[DTI05]		✓			✓				✓	
4.	[Mar05]		✓		✓			✓			
5.	[LV05]		✓			✓		✓			
6.	[AW06]		✓		✓	✓		✓	✓		
7.	[EHHJ07]		✓		✓			✓	✓	✓	
8.	[BKM08]	✓	✓		✓				✓		
9.	[BA11]		✓		✓			✓	✓		✓
10.	[ZZZLy12]		✓		✓			✓			✓
11.	[STWJ13]			✓		✓	✓				

Tabelle 3.2.: Charakteristiken einiger generalisierter modellbasierter Filterfunktionen.

Amplituden Gebrauch machen, verwenden die meisten von ihnen die generalisierten Verteilungen. Eine besondere Stellung genießt hier allerdings die Filterfunktion aus [BKM08], welche sowohl die generalisierten SA als auch die generalisierten Verteilungen verwendet. Während die meisten Beiträge sich mit den MMSE-Schätzern beschäftigen, gibt es auch eine Reihe von MAP-Schätzern. Von den vier Verteilungen, die bei den generalisierten Schätzern eingesetzt werden, umfasst die dreiparametrige generalisierte Gamma-Verteilung (GGV) die Rayleigh-, die Gamma- und die Chi-Verteilung [KA10]. In ihrer allgemeinen Form wird sie eigentlich nur in [DTI05] verwendet, denn in [EHHJ07] wird sie nur eingeführt. Die in [EHHJ07] hergeleiteten Schätzer basieren auf den beiden Spezialfällen der generalisierten Gamma-Verteilung, der Chi- und der Gamma-Verteilung. Die letzten beiden sind zweiparametrig und dementsprechend handhabbarer für die Berechnungen[9]. Aber auch die Rayleigh-Verteilung wird für die statistische Modellierung von den Schätzern verwendet.

3.3. Verfahren zur *a priori* SNR-Schätzung

Soweit dem Autor bekannt, wird das *a priori* SNR in der spektralen Sprachsignalentstörung zum ersten Mal in [MM80] eingeführt, wo es als eine Zufallsvariable $|S(k,\ell)|^2/\lambda_D(k,\ell)$ definiert wird, also ohne den Erwartungswertoperator im Zähler. Die gebräuchliche Definition des *a priori* SNR aus (2.22) findet man erst in [EM83]. Hier ist $\xi(k,\ell)$ ein Parameter der Verteilungsdichtefunktion des *a posteriori* SNR $\gamma(k,\ell)$, das als eine Zufallsvariable definiert wird. Unter den Modellannahmen (2.4) und (2.5) für $S(k,\ell)$ und $D(k,\ell)$ mit der Additivität (2.3) ist $\gamma(k,\ell)$ eine exponentialverteilte Zufallsvariable, in dessen Skalierungsparameter $\xi(k,\ell)$ einfließt [Coh03]. Aus diesem Grund ist es intuitiv, $\xi(k,\ell)$ aus $\gamma(k,\ell)$ direkt im SNR-Bereich schätzen zu wollen. Eine Möglichkeit dafür bietet der ML-Schätzwert

$$\hat{\xi}^{\mathrm{ML}}(k,\ell) = \max(\hat{\gamma}(k,\ell) - 1,\, 0), \tag{3.12}$$

wobei ein Maximumoperator $\max(\cdot)$ sicherstellt, dass $\hat{\xi}^{\mathrm{ML}}(k,\ell)$ positiv ist [EM84]. Zu beachten ist, dass in (3.12) statt $\gamma(k,\ell)$ sein Schätzwert $\hat{\gamma}(k,\ell)$ eingesetzt wird, der laut (2.16) mit Hilfe des RLDS-Schätzwertes $\hat{\lambda}_D(k,\ell)$ berechnet wird. Der ML-Schätzer (3.12) ist zwar einfach, bringt jedoch eine unverhältnismäßig hohe Schätzfehlervarianz mit sich. Aus diesem Grund wird in [EM84] vorgeschlagen, den ML-Schätzer in eine Rekursionsgleichung erster Ordnung wie (2.48) mit einer Glättungskonstante $\alpha_{\mathrm{ML}} \in (0;\, 1)$ zu integrieren, und so die Schätzfehlervarianz zu reduzieren. Allerdings scheitert dieser Ansatz in einer Realisierung, denn die Verwendung einer relativ großen Glättungskonstante $(1 - \alpha_{\mathrm{ML}}) \to 0$, die für eine ausreichende Reduzierung der Schätzfehlervarianz notwendig ist, zu einer inakzeptablen Verzögerung der *a priori* SNR-Schätzwerte führt, die sich besonders bei Energiesprüngen des Sprachsignals bemerkbar macht und zum unerwünschten *musical noise* beiträgt.

Decision-Directed Verfahren: Aus diesem Grund wird in [EM84] das DD-Verfahren entwickelt, das eine Bildung von *musical noise* weitestgehend vermeidet. Das DD-Verfahren ist der mit Abstand am weitesten verbreitete *a priori* SNR-Schätzer. Seine Recheneffizienz trägt nicht zuletzt zu seiner hohen Akzeptanz bei. Berechnet wird der DD-Schätzwert als

[9] Da sowohl die Gamma- als auch die Chi-Verteilung einen Formparameter haben, werden die aus diesen Modellen resultierenden Schätzer als generalisierte Schätzer betrachtet und zwar ungeachtet dessen, dass in den Namen diese beiden Verteilungen das Wort 'generalisiert' fehlt.

eine gewichtete Summe zweier gewichteter Terme

$$\hat{\xi}^{\mathrm{DD}}(k,\ell) = \alpha_{\mathrm{DD}} \cdot \tilde{\xi}(k,\ell-1) + (1-\alpha_{\mathrm{DD}}) \cdot \hat{\xi}^{\mathrm{ML}}(k,\ell), \tag{3.13}$$

wobei α_{DD} ein Gewichtsfaktor ist und $\tilde{\xi}(k,\ell-1)$ sich als ein vom vergangenen Rahmen propagierter *a priori* SNR-Schätzwert darstellt

$$\tilde{\xi}(k,\ell-1) = \frac{|\hat{S}(k,\ell-1)|^2}{\hat{\lambda}_D(k,\ell-1)} = G^2(k,\ell-1) \cdot \hat{\gamma}(k,\ell-1), \tag{3.14}$$

der im Weiteren auch als Propagationsterm bezeichnet wird. Dabei fällt auf, dass der Propagationsterm $\tilde{\xi}(k,\ell-1)$ ähnlich wie in [MM80] definiert wird, jedoch mit dem Unterschied, dass statt der wahren Amplitude $|S(k,\ell-1)|$ der Schätzwert $|\hat{S}(k,\ell-1)|$ verwendet wird. Außerdem macht das DD-Verfahren vom ML-Schätzer aus (3.12) Gebrauch, der im aktuellen Rahmen berechnet wird und als zweiter Summand in (3.13) agiert. Somit kombiniert das DD-Verfahren die zwei Definitionen aus [MM80] und [EM84] und lässt sie in einem gemeinsamen Schätzwert verschmelzen. Man beachte, dass obwohl (3.13) eine große Ähnlichkeit mit der rekursiven Glättung (2.48) aufweist, ist sie keine Glättung an sich, denn es gilt im allgemeinen $\tilde{\xi}(k,\ell-1) \neq \hat{\xi}^{\mathrm{DD}}(k,\ell-1)$. Aus diesem Grund wird α_{DD} in [EM84] explizit nicht als eine Glättungskonstante sondern als ein Gewichtsfaktor bezeichnet, dessen Wahl die Leistungsfähigkeit des Schätzers stark beeinflusst. Die typischen Werte von α_{DD} werden häufig im Bereich $[0.92;\ 0.99]$ relativ groß gewählt [BM10].

Aus der Definition (3.14) wird ersichtlich, dass das DD-Verfahren auf einer engen Zusammenarbeit von den Bausteinen 1 und 2 des Systems aus Abb. 2.2 aufbaut, was ihn zum Dreh- und Angelpunkt im System macht. Wie in [Coh01] berechtigt erwähnt, wird das DD-Verfahren unter Annahme der Sprachsignalpräsenz hergeleitet. Daraus folgt, dass für die Berechnung von $\tilde{\xi}(k,\ell-1)$ statt der Filterfunktion $G(k,\ell-1)$ aus (2.19) oder (2.20), die für Signalentstörung verwendet wird, die Filterfunktion $G_{H_1}(k,\ell-1)$ einzusetzen ist, die für die Sprachsignalanwesenheit hergeleitet wird. In [EM84] werden dabei drei Filterfunktionen betrachtet: das Wiener-Filter, der MMSE-SA-Schätzer und die Filterfunktion aus [MM80]. In Abhängigkeit von der verwendeten Filterfunktion soll auch ein bestimmter Wert des Gewichtsfaktors α_{DD} verwendet werden. In [Yu13] wird ein konstanter Gewichtsfaktor für den MMSE-LSA-Schätzer aus [EM85] und für die Filterfunktion aus [LV05] experimentell optimiert. Aus der Optimierung wird noch einmal deutlich, dass der Wert von α_{DD} zweckgebunden ist und unterschiedlich gewählt werden kann, je nachdem, ob das entstörte Signal entweder die beste Qualität oder die stärkste Störsignaldämpfung aufweisen soll.

Allerdings weist der DD-Schätzer einen bekannten Nachteil auf, nämlich einen Nachhalleffekt (engl. *reverberation effect*), der sich als eine langsame Verfolgung einer plötzlichen Änderung des momentanen SNR-Niveaus darstellt [PMS06]. Der Hauptgrund dafür ist der konstante Gewichtsfaktor α_{DD}, der ja relativ groß gewählt werden muss, um *musical tones* im prozessierten Signal zu vermeiden. Dies war ein Anlass für viele Forscher, das DD-Verfahren näher zu untersuchen und darauf basierend auch weiter zu entwickeln. So führte [Cap94] eine unteren Schranke ξ_{min} für $\hat{\xi}^{\mathrm{DD}}(k,\ell)$ ein, wie bereits in Abschnitt 2.2 ausführlich diskutiert wurde. Tiefgreifende Untersuchungen des DD-Verfahrens in [BM10] zeigen einige weitere Besonderheiten des Schätzers z. B. bei seiner Verwendung in Kombination mit der MMSE-LSA-Filterfunktion aus [EM85]. In diversen Weiterentwicklungen wird das DD-Verfahren entweder hinsichtlich der Modifikation des Gewichtsfaktors α_{DD} oder der Umgestaltung des ersten Summanden $\tilde{\xi}(k,\ell-1)$ optimiert.

Modifizierte DD-Verfahren mit einem zeitinvarianten Gewichtsfaktor: In [SK00] wird vorgeschlagen statt eines konstanten Gewichtsfaktors α_{DD} einen zeitvarianten frequenzunabhängigen Gewichtsfaktor $\alpha_{\mathrm{DD}}(\ell)$ zu verwenden, welcher mittels der momentanen Leistungsänderung der aufeinander folgenden Rahmen heuristisch adaptiert wird. In [HSK04] wird ein frequenzabhängiger zeitvarianter Gewichtsfaktor vorgeschlagen, der basierend auf einer statistischen Modellierung im MMSE-Sinne optimal gewählt wird

$$\alpha_{\mathrm{DD}}(k,\ell) = \frac{1}{1+\left(\frac{\hat{\xi}^{\mathrm{ML}}(k,\ell)-\tilde{\xi}(k,\ell-1)}{\hat{\xi}^{\mathrm{ML}}(k,\ell)+1}\right)^2}.$$

Wie experimentelle Untersuchungen zeigen, gelingt es mit dieser Technik, das mittlere segmentelle SNR leicht zu verbessern. Mit der Adaption des Gewichtsfaktors mittels einer Sigmoid-Funktion, die von der Differenz der aufeinander folgenden *a posteriori* SNR-Schätzwerte abhängt, beschäftigt sich [PC07a]. In den Experimenten führt diese Technik verglichen mit dem DD-Verfahren zur besseren Qualität der prozessierten Signale. In [YZZ16] wird der Gewichtsfaktor $\alpha_{\mathrm{DD}}(k,\ell)$ über die Veränderung der lokalen Entropie gesteuert, die aus den geglätteten spektralen Amplituden des gestörten Sprachsignals der fünf vergangenen STFT-Rahmen berechnet wird. Während eine kleine Entropieveränderung mit der Sprachsignalabwesenheit einhergeht und zur Wahl des größeren Gewichtsfaktors führt, ist eine große Entropieveränderung ein Indikator für die Sprachsignalpräsenz, bei der der Gewichtsfaktor gesenkt wird. Verglichen mit dem konventionellen DD-Verfahren werden in [YZZ16] prozessierte Sprachsignale mit einem höheren segmentellen SNR betrachtet.

Weitere modifizierte DD-Verfahren: In [PMS06] wird ein zweistufiges DD-Verfahren entwickelt. Im ersten Schritt wird hier mit dem konventionellen DD-Verfahren zunächst die Filterfunktion $G_{\mathrm{DD}}(k,\ell)$ im aktuellen Zeit-Frequenz-Punkt berechnet. Im zweiten Schritt wird diese für die Berechnung des finalen *a priori* SNR-Schätzwertes als $\frac{|G_{\mathrm{DD}}(k,\ell)\cdot Y(k,\ell)|^2}{\hat{\lambda}_D(k,\ell)}$ verwendet. In [YND13] wird vorgeschlagen, die im $\tilde{\xi}(k,\ell-1)$ beteiligten Größen möglichst aus dem aktuellen Rahmen zu verwenden. Und da zum Zeitpunkt der Berechnung des *a priori* SNR-Schätzwertes nur der Wert der aktuellen Filterfunktion unbekannt ist, wird in (3.14) der *a posteriori* SNR-Schätzwert $\hat{\gamma}(k,\ell-1)$ einfach durch $\hat{\gamma}(k,\ell)$ ersetzt. Allerdings macht diese Modifikation eine zusätzliche Glättung der *a posteriori* SNR-Schätzwerte erforderlich, um *musical noise* weiterhin vermeiden zu können. Eine andere Erweiterung des DD-Verfahrens wird in [SOLG14] vorgeschlagen. Der Gleichung (3.13) fügt man hier einen zusätzlichen Schwungterm (engl. *momentum term*) hinzu, dessen Gewicht mit einem zeitvarianten Vorfaktor adaptiert wird, der im MMSE-Sinne optimal gewählt wird. In [LLC14] wird die Gleichung (3.13) als ein Regressionsmodell betrachtet, dessen Parameter im Rahmen einer *Least Squares* Optimierung berechnet werden. Ein Vergleich mit dem DD-Verfahren und dem antikausalen *a priori* SNR-Schätzer aus [Coh05] zeigt, dass der Schätzer aus [LLC14] die beiden Verfahren hinsichtlich der Qualität der prozessierten Sprachsignale übertrifft.

Alternative Schätzverfahren: Mit der Zeit wurden auch einige alternativen Verfahren zur *a priori* SNR-Schätzung entwickelt. So werden in [Coh05] zwei *a priori* SNR-Schätzer vorgestellt, welche die zeitlichen Korrelationen zwischen den benachbarten spektralen Komponenten eines Sprachsignals berücksichtigen. Die beiden Schätzer beinhalten solche Prozessierungsschritte wie ein Propagationsschritt und ein Aktualisierungsschritt, die aus der konventionellen Kalman-Filterung bekannt sind. Während ein Schätzer kausal ist, stellt sich der andere als ein antikausaler Schätzer dar, der auf den Beobachtungen einiger weniger

Rahmen arbeitet, die sich relativ zum aktuellen Rahmen in der unmittelbaren Zukunft befinden. Dadurch werden *a priori* SNR-Schätzwerte des antikausalen Verfahrens mit einer kleinen Verzögerung berechnet, die in manchen Anwendungen jedoch toleriert werden kann. Im Unterschied zum kausalen Schätzverfahren kann der antikausale Schätzer zwischen den Änderungen in der momentanen spektralen Leistung unterscheiden, die durch das Sprachsignal wie *speech onsets* oder durch das nichtstationäre Störsignal hervorgerufen werden. In Experimenten mit den stark gestörten Sprachsignalen übertrifft der antikausale *a priori* SNR-Schätzer sowohl das DD-Verfahren als auch den kausalen *a priori* SNR-Schätzer deutlich. Der kausale *a priori* SNR-Schätzer, der in einem Spezialfall zum DD-Verfahren mit einem zeitvarianten frequenzabhängigen Gewichtsfaktor wird, schneidet in den Experimenten in etwa wie das konventionelle DD-Verfahren mit einem konstanten Gewichtsfaktor ab.

Obwohl die meisten Schätzverfahren den *a priori* SNR-Schätzwert $\hat{\xi}(k,\ell)$ aus dem Schätzwert des *a posteriori* SNR $\hat{\gamma}(k,\ell)$ berechnen, ist dies nicht die einzige Möglichkeit. Laut der Definition (2.22) kann man $\hat{\xi}(k,\ell)$ auch berechnen, wenn neben der Schätzung der Rauschleistungsdichte $\lambda_D(k,\ell)$ auch das Leistungsdichtespektrum des ungestörten Sprachsignals $\lambda_S(k,\ell)$ geschätzt wird. So wird $\hat{\lambda}_S(k,\ell)$ in [BGM08] im Bereich von sogenannten Cepstren über die rekursive Glättung des ML-Schätzwertes $\hat{\lambda}_S^{\mathrm{ML}}(k,\ell)$ berechnet, der aus der ML-Schätzung $\hat{\xi}^{\mathrm{ML}}(k,\ell)$ aus (3.12) bestimmt wird. Dabei wird eine zeitvariante Glättungskonstante verwendet, die mit Hilfe der geschätzten Grundfrequenz der Stimmbänder

	Quelle	ML-Schätzer	MAP-Schätzer	klassischer Propagationsterm	geänderter Propagationsterm	konstanter Gewichtsfaktor	zeitvarianter Gewichtsfaktor	klassische Gewichtung	erweiterte Gewichtung	Alternatives Verfahren
1.	[EM84]	✓		✓		✓		✓		
2.	[SK00]	✓		✓			✓	✓		
3.	[HSK04]	✓		✓			✓	✓		
4.	[Coh05]	✓		✓			✓		✓	✓
5.	[PC07a]	✓		✓			✓	✓		
6.	[PMS06]	✓			✓	✓		✓		
7.	[BGM08]	✓								✓
8.	[YND13]	✓			✓	✓		✓		
9.	[LLC14]				✓	✓			✓	✓
10.	[SOLG14]	✓		✓		✓			✓	
11.	[EMTF15]	✓								✓
12.	[YZZ16]	✓		✓			✓	✓		
13.	[CHHU16]		✓							✓

Tabelle 3.3.: Eigenschaften einiger Verfahren zu *a priori* SNR-Schätzung.

(engl. *fundamental frequency*) adaptiert wird. Das Verfahren erweist sich leistungsfähiger als das DD-Verfahren, benötigt allerdings einige zusätzlichen Parameter, die heuristisch gewählt werden. Für die Berechnung von $\hat{\lambda}_S(k,\ell)$ in [EMTF15] werden die spektralen Einhüllenden $(\,|S(0)|,|S(1)|,\ldots,|S(k)|,\ldots,|S(k_{\text{Nyq}})|)^{\text{T}}$ mit einem multivariaten Gaußschen Mischungsmodell mit 512 Mischungskomponenten statistisch modelliert, dessen Gewichte, Mittelwertvektoren und diagonale Kovarianzmatrizen im Rahmen einer Trainingsphase bestimmt werden. Beim Vergleich mit dem DD-Verfahren wird wird hier eine Reduzierung des *a priori* SNR-Schätzfehlers in der Sprachsignalanwesenheit beobachtet, was zu prozessierten Sprachsignalen mit derselben Qualität und besserer Rauschunterdrückung führt. Sonst bleibt noch zu erwähnen, dass in [CHHU16] ein MAP-basierter Schätzer von *a priori* SNR unter Verwendung von Weibull-Mischungsmodellen entwickelt wird. Dabei werden Vorteile einer statistischen Modellierung im Bereich der generalisierten spektralen Amplituden ausgenutzt.

Einige wichtige Eigenschaften der diskutierten *a priori* SNR-Schätzer sind in Tab. 3.3 zusammengefasst. Obwohl es eine breite Palette an verschiedensten *a priori* SNR-Schätzern gibt, bleibt das DD-Verfahren immer noch sehr weit verbreitet und beliebt, wie die Analyse der Fachliteratur zur spektralen Sprachsignalentstörung zeigte.

3.4. Berechnung der Sprachpräsenzwahrscheinlichkeit

Die STFT-Koeffizienten benachbarter Zeit-Frequenz-Punkte eines Sprachsignals sind weder entlang der Zeit- noch entlang der Frequenzachse unkorreliert. Während zeitliche Korrelationen mit einer kurzzeitigen Stationarität eines Sprachsignals einhergehen, hängen Korrelationen entlang der Frequenzachse mit der harmonischen Struktur eines Sprachsignals zusammen [MHA14]. Die Ausnutzung dieser ZF-Korrelationen gehört zu den wichtigen Standardtechniken moderner SPP-Schätzer, die in der spektralen Sprachsignalentstörung eingesetzt werden. Nicht selten werden auch Systeme entwickelt, in denen sogar zwei unterschiedliche SPP-Schätzer mit verschiedenen Zielsetzungen realisiert werden [CB01b]. So gehört eine VAD/SPP Schätzung zu den wichtigsten Techniken moderner RLDS-Schätzer, wie in Abschnitt 3.1 bereits erwähnt wurde. In diesem Abschnitt allerdings werden modellbasierte Verfahren zur SPP-Schätzung vorgestellt, welche in die Berechnung spektraler Filterfunktion einfließen und somit laut Abb. 2.2 als vierter Baustein eines Systems zur spektralen Sprachsignalentstörung zum Einsatz kommen. Einige Aspekte zur Motivation, zur konventionellen statistischen Modellierung und zu den Einsatzmöglichkeiten solcher SPP-Schätzer in der Sprachsignalentstörung wurden am Ende von Abschnitt 2.2 bereits kurz vorgestellt.

Schätzverfahren ohne explizite Berücksichtigung von ZF-Korrelationen: In [MM80] wird eine ML-basierte spektralen Filterfunktion mit einer SPP-Schätzung multiplikativ (wie in (2.38) mit $G_{H_0} = 0$) überlagert. Diese Überlagerung führt zu einer ML-basierten *soft-decision* Filterfunktion, die eine stärkere Dämpfung in den energieschwachen Zeit-Frequenz-Punkten ausübt und somit zur stärkeren Störsignaldämpfung führt. Da die verwendete SPP-Schätzung ohne Berücksichtigung der Zeit-Frequenz-Korrelationen berechnet wird, wird die resultierende Filterfunktion einer zusätzlichen rekursivartigen Glättung unterzogen, um sprunghafte Veränderungen der Filterfunktion zu reduzieren und somit *musical tones* möglichst zu vermeiden. In den subjektiven Hörexperimenten führte die ML-basierte Filterfunktion kombiniert mit einer SPP-Schätzung zur höheren Rauschunterdrückung allerdings auf Kosten von Verzerrungen der Sprachsignalkomponenten.

In [EM84] wird inspiriert vom MMSE-Schätzer aus [ME68] eine MMSE-SA-Filterfunktion aus [EM83] mit einer SPP-Schätzung multiplikativ überlagert. Während der SPP-Schätzer hier ebenso wie in [MM80] den zweiten Summanden aus (2.38) vernachlässigt und keine Zeit-Frequenz-Korrelationen explizit ausnutzt, wird das *a priori* SNR mit dem vorgeschlagenen *Decision-Directed* Verfahren geschätzt, das sich in gewissem Sinne als eine zeitliche Glättung darstellt. Ein subjektiver Vergleich der prozessierten Sprachsignale ergibt, dass der vorgeschlagene SPP-basierte MMSE-SA-Schätzer verglichen mit der MMSE-SA-Filterfunktion ohne einer SPP-Schätzung zu einer stärkeren Reduktion des Restrauschens und gegenüber dem Schätzer aus [MM80] zu den prozessierten Sprachsignalen mit weniger *musical tones* führt.

Verfahren zur SPP-Schätzung einmal pro STFT-Rahmen: In [Yan93] wird vorgeschlagen G_{H_0} aus (2.38) auf einen minimalen Wert der spektralen Filterfunktion zu setzen. Man beachte, dass in [MM80] und [EM84] der zweite Summand aus (2.38) durch Verwendung von $G_{H_0} = 0$ unberücksichtigt blieb. Außerdem wird $\mathcal{P}(k, \ell)$ in [Yan93] einmal pro STFT-Rahmen geschätzt. Dafür wird zunächst ein mittleres momentanes SNR im aktuellen Rahmen berechnet, das anschließend mit einer zuvor festgelegten Schwelle verglichen wird, um eine binäre Entscheidung zu treffen. Verwendet man die ML-basierte spektrale Filterfunktion aus [MM80] als $G_{H_1}(k, \ell)$ und den vorgeschlagenen rahmenweisen VAD-Schätzer als $\mathcal{P}(k, \ell)$, resultiert daraus eine modifizierte ML-basierte spektrale Filterfunktion. Verglichen mit fünf alternativen spektralen Filterfunktionen zeigt die vorgeschlagene Entstörungsregel in den subjektiven und objektiven Tests die beste Leistungsfähigkeit.

In [SKS99] wird ein erwähnenswertes Verfahren zur SPP-Schätzung präsentiert, das die Zeit-Frequenz-Korrelationen mittels Hidden-Markov-Modelle (HMM) modelliert und dadurch bei der Schätzung ausnutzt. Dabei wird ähnlich wie in [Yan93] eine harte VAD-Entscheidung für einen ganzen STFT-Rahmen getroffen, ob in diesem Rahmen Sprachsignalpräsenz detektiert wird oder nicht. Für eine Schätzung des *a priori* SNR, das in die Entscheidungsregel miteinfließt, wird dabei entweder der ML-Schätzer aus (3.12) oder das konventionelle DD-Verfahren aus (3.13) verwendet. Das charakteristische Merkmal dieses SPP-Schätzers ist jedoch die Verwendung des Hidden-Markov-Modells, das sowohl Korrelationen entlang der Zeitachse als auch entlang der Frequenzachse modelliert. Während frequenzübergreifende Korrelationen dadurch berücksichtigt werden, dass eine gemeinsame Entscheidung über Beobachtungen aller Frequenzbänder getroffen wird, werden zeitliche Korrelationen durch die HMM-Transitionswahrscheinlichkeiten mit einbezogen. In den Experimenten mit den Sprachsignalen führt die Verwendung des HMM-basierten SPP-Schätzers zur konsistenten Erhöhung der Detektionswahrscheinlichkeit für eine gegebene Falschalarmwahrscheinlichkeit unabhängig von der Art des verwendeten *a priori* SNR-Schätzers.

Auch in [KC00] wird die Sprachpräsenzwahrscheinlichkeit einmal in einem ganzen STFT-Rahmen geschätzt. Die Besonderheit hier ist die Art und Weise, wie die geschätzte SPP in die Entstörungsregel einfließt. Statt eine spektrale Filterfunktion multiplikativ oder mittels einer Potenzierung zu modifizieren, werden Schätzwerte von *a posteriori* SNR $\hat{\gamma}(k, \ell)$ und von *a priori* SNR $\hat{\xi}(k, \ell)$ mit den SPP-Schätzwerten multipliziert. Anschließend werden die resultierenden modifizierten SNR-Größen bei der Berechnung der Filterfunktion aus [EM84] verwendet. Sonst wird hier eine besonders hohe *a priori* SAP von $q \approx 0.94$ verwendet, welche durch die gut bekannte Dünnbesetztheit von Sprachsignalen gerechtfertigt werden kann. Die subjektiven Hörtests zeigten die Überlegenheit des vorgeschlagenen Verfahrens gegenüber einem konventionellen Entstörungsalgorithmen.

Schätzverfahren mit zeitvarianter *a priori* SAP: Während in vielen Verfahren die *a priori* Wahrscheinlichkeit für die Sprachsignalabwesenheit $q = \Pr(H_0)$ aus (2.37) als eine konstante Größe betrachtet wird, wird sie in [MCA99] als eine zeitvariante frequenzabhängige Variable $q(k, \ell)$ aufgefasst und im Rahmen des vorgeschlagenen Verfahrens geschätzt. Außerdem ist in dieser Publikation Verwendung der spektralen MMSE-LSA-Filterfunktion $G_{\text{LSA}}(k, \ell)$ aus [EM85] etwas ungewöhnlich, bei welcher die Sprachpräsenzwahrscheinlichkeit $\mathcal{P}(k, \ell)$ in die Berechnung der finalen Filterfunktion (2.39) in der Regel als Potenz eingeht. In [MCA99] wird sie jedoch ähnlich wie in [MM80, EM84] als Vorfaktor verwendet, woraus eine *multiplicatively-modified* LSA (MM-LSA) Filterfunktion entsteht. Für die Schätzung von SAP $q(k, \ell)$ wird hier zunächst ein Neyman-Pearson-Kriterium verwendet, das auf den *a posteriori* SNR-Schätzwerten $\hat{\gamma}(k, \ell)$ mit Hilfe einer harten Schwellenwertentscheidung realisiert wird. Die daraus resultierenden binären Indikatoren $I(k, \ell)$ werden anschließend in einer Rekursionsgleichung mit einem konstanten Glättungsfaktor über die Zeit geglättet, um SAP-Schätzwerte $\hat{q}(k, \ell)$ zu berechnen. In einer subjektiven Bewertung der Qualität der prozessierten Sprachsignale erreichte die MM-LSA-Filterfunktion kombiniert mit einem moderaten Wert G_{H_0} gute Ergebnisse.

In [SKY99] wird die SPP-basierte MMSE-SA-Filterfunktion aus [EM84] mit zwei unterschiedlichen SAP-Schätzern eingesetzt. Die erste SAP-Schätzung wird ähnlich wie in [MCA99] berechnet mit dem einzigen Unterschied, dass hier statt Neyman-Pearson-Kriterium eine ML-Entscheidungsregel verwendet wird, die keine heuristische Entscheidungsschwelle benötigt. Der zweite SAP-Schätzer wird mit Hilfe der Bayesschen Regel für bedingte Wahrscheinlichkeiten hergeleitet, bei der allerdings angenommen wird, dass die *a priori* SAP gleich 0.5 sein muss. Diese Annahme führt zu einer Inkonsistenz in der Herleitung des zweiten SAP-Schätzers. In einer objektiven experimentellen Untersuchung übertreffen beide SAP-Schätzer eingesetzt in der SPP-basierten MMSE-SA-Filterfunktion mit einem konstanten *a priori* SAP $q = 0.2$ den ursprünglichen Schätzer aus [EM84], erreichen jedoch in etwa eine ähnliche Leistungsfähigkeit mit leichten Vorteilen für den zweiten Schätzer.

Auch in [Coh01] wird vorgeschlagen, die SAP als eine zeitvariante frequenzabhängige Größe zu betrachten. Für ihre Schätzung in einem Zeit-Frequenz-Punkt werden hier zum ersten Mal drei verschiedene getrennt definierte Zeit-Frequenz-Umgebungen um den (k, ℓ)-ten Punkt betrachtet. Neben der Schätzung einer rahmenweisen SPP-$\mathcal{P}_{\text{frame}}(\ell)$ werden außerdem eine lokale und eine globale Umgebungen definiert, die jeweils für eine Schätzung von $\mathcal{P}_{\text{loc}}(k, \ell)$ und $\mathcal{P}_{\text{glob}}(k, \ell)$ benutzt werden. Dabei wird die resultierende SAP als $q(k, \ell) = 1 - \mathcal{P}_{\text{loc}}(k, \ell) \cdot \mathcal{P}_{\text{glob}}(k, \ell) \cdot \mathcal{P}_{\text{frame}}(\ell)$ berechnet. Als Berechnungsgrundlage aller drei SPP-Wahrscheinlichkeiten dienen dabei *a priori* SNR-Schätzwerte, die zunächst über die Zeit rekursiv geglättet werden. Während die im ℓ-ten Rahmen berechneten geglätteten *a priori* SNR-Schätzwerte gemittelt über alle Frequenzbänder in die Ermittlung von $\mathcal{P}_{\text{frame}}(\ell)$ eingehen, fließt nur ihre begrenzte Anzahl in die Berechnung von $\mathcal{P}_{\text{loc}}(k, \ell)$ und $\mathcal{P}_{\text{glob}}(k, \ell)$ ein[10]. Wie Untersuchungen in [Coh01] zeigten, trägt eine gleichzeitige Verwendung verschiedener Umgebungen zur Reduzierung der Außreißeranzahl in der endgültigen SAP-Schätzung bei. Außerdem wird hier die bereits erwähnte OMLSA-Filterfunktion hergeleitet, in welche die SPP-Schätzung als eine Potenz eingeht. In einem objektiven expe-

[10] Während bei der Berechnung von $\mathcal{P}_{\text{loc}}(k, \ell)$ geglättete *a priori* SNR-Schätzwerte der beiden direkt benachbarten Frequenzbänder $k - 1$ und $k + 1$ verwendet werden, werden für die Berechnung von $\mathcal{P}_{\text{glob}}(k, \ell)$ in [Coh01] geglättete *a priori* SNR-Schätzwerte der 15 benachbarten Frequenzbänder (also insgesamt 31 Werte) benutzt, die zuvor mit einem normalisierten Hanning-Fenster gewichtet werden.

rimentellen Vergleich mit den Filterfunktionen aus [EM84, EM85, MCA99] erreichte der vorgeschlagene OMLSA-Schätzer die größten Werte des segmentellen SNR.

Verfahren mit Verwendung einer lokalen und einer globalen Schätzung: Das Schätzverfahren aus [SA05] verfolgt das Ziel, zu einer SPP-Schätzung mit den sogenannten verbundenen Zeit-Frequenz-Regionen (engl. *connected time-frequency speech presence regions*) zu gelangen. Dafür werden zwei binäre VAD-Schätzwerte im Zeit-Frequenz-Bereich berechnet, die ähnlich wie in [Coh01] für eine finale VAD-Schätzung miteinander multipliziert werden. Entscheidungsschwellen der beiden VAD-Schätzer werden dabei so ausgelegt, dass eine der VAD-Entscheidungen mit einer lokalen SPP $\mathcal{P}_{\text{loc}}(k,\ell)$ und die andere mit einer globalen SPP $\mathcal{P}_{\text{glob}}(k,\ell)$ verglichen werden kann. Dabei werden beide Entscheidungen basierend auf einem geglätteten Spektrogramm getroffen, das aus dem Spektrogramm des gestörten Sprachsignals mittels einer Glättung berechnet wird, die sowohl entlang der Frequenz als auch entlang der Zeit stattfindet. Als Entstörungsregel wird die generalisierte spektrale Subtraktion wie in (3.4) aus [BSM79] mit dem Kompressionsfaktor $\beta = 0.8$ verwendet, die allerdings nur dann angewandt wird, wenn die VAD-Schätzung die Anwesenheit eines Sprachsignals signalisiert. Da die verwendete Entstörungsregel weder Gebrauch von *a priori* SAP noch von *a priori* SNR macht, werden diese im Rahmen dieses Verfahrens auch nicht benötigt. In einem experimentellen Vergleich führt die vorgeschlagene VAD-basierte Filterfunktion zu prozessierten Signalen mit einer besseren Qualität als der MMSE-LSA-Schätzer aus [EM85], jedoch unterliegt sie diesem hinsichtlich der Störsignaldämpfung.

In [GBM08] werden sowohl eine lokale als auch eine globale Zeit-Frequenz-Umgebung mit einer fest definierten Anzahl an verwendeten Zeit-Frequenz-Punkten verwendet. Man beachte, dass die ZF-Umgebungen bei den Verfahren aus [Coh01] und [SA05] durch die Verwendung einer rekursiven Glättung entlang der Zeitachse eigentlich unbegrenzt sind. Außerdem wird in [GBM08] vorgeschlagen, für eine SPP-Berechnung einen fixierten *a priori* SNR-Wert zu verwenden, der im Rahmen einer Optimierungsaufgabe gefunden und während einer Signalverarbeitung nicht mehr geschätzt wird. In diesem Fall gibt es keine Kopplung einer SPP-Schätzung mit einer *a priori* SNR-Schätzung, die sich sonst in der Sprachsignalabwesenheit negativ auf Leistungsfähigkeit eines SPP-Schätzers auswirken kann [MWJ00]. Außerdem wird hier auch keine SAP $q(k,\ell)$ geschätzt, sondern auf ihren *a posteriori* SAP-Wert konstant gesetzt. Die Verteilungsdichten $p_{\gamma(k,\ell)|H_1}(\gamma)$ und $p_{\gamma(k,\ell)|H_0}$ werden mit einer skalierten Chi-Quadrat-Verteilung aus [ML01] modelliert. Durch eine experimentelle Anpassung eines der Modellparameter werden die STFT-bedingte Korrelationen zwischen den Beobachtungen einzelner Zeit-Frequenz-Punkte mitberücksichtigt. In einer experimentellen Untersuchung werden SPP-Schätzer aus [MCA99] und [Coh01] vom vorgeschlagenen Verfahren übertroffen, das einen besseren Kompromiss zwischen Störsignaldämpfung und Sprachsignalqualität findet. Man beachte, dass das Verfahren aus [GBM08] als Ausgangspunkt für Entwicklung einiger anderen SPP-Schätzer diente.

Verfahren mit generalisierten spektralen Filterfunktionen: Auch bei den generalisierten spektralen Filterfunktionen kommt eine SPP-Schätzung zum Einsatz. Einer der ersten solchen Schätzer hier ist das Verfahren aus [DA09], in dem die spektrale MMSE-GSS-Filterfunktion (3.11) aus [YKR05] verwendet wird. Der SPP-Schätzer aus [Coh01] wird hier auf die Filterfunktion multiplikativ wie in (2.38) angewandt, jedoch direkt im Bereich der generalisierten spektralen Amplituden. Aus diesem Grund weist der hier vorgeschlagene SPP-Schätzer die gleichen Eigenschaften wie der Schätzer aus [Coh01] auf. Außerdem wird anders als in [YKR05] vorgeschlagen, den adaptiven Kompressionsfaktor $\beta(k,\ell)$ über

eine lineare Abhängigkeit vom verwendeten SPP-Schätzer zu berechnen. Inspiriert durch die Untersuchungen in [PC07b] nimmt der Kompressionsfaktor $\beta(k, \ell)$ dabei nur die negativen Werte aus dem Bereich $[-0.8; 0]$ an. In den objektiven experimentellen Untersuchungen erreicht der vorgeschlagene SPP-basierte generalisierte Schätzer bessere Leistungsfähigkeit als der OMLSA-Schätzer aus [CB01b] und der MMSE-GSS-Schätzer aus [YKR05].

In [FF12] wird die generalisierte MMSE-basierte spektrale Filterfunktion ohne SPP-Schätzung aus [EHHJ07] verwendet, bei welcher die spektralen Amplituden eines Sprachsignals mit der generalisierten Gamma Verteilungsdichtefunktion modelliert werden. Im Unterschied dazu werden bei Herleitung des vorgeschlagenen SPP-Schätzers, der die Filterfunktion multiplikativ überlagert, komplexwertige STFT-Koeffizienten eines Sprachsignals mit einer bivariaten generalisierten Verteilungsdichtefunktion modelliert. Um zu einer realisierbaren SPP-basierten Filterfunktion zu gelangen, muss allerdings einer der Formparameter der verwendeten generalisierten Verteilungen ähnlich wie in [LV05] fixiert werden. Eine objektive Experimentenreihe zeigte, dass die vorgeschlagene Filterfunktion im Vergleich zum SPP-basierten MMSE-Schätzer aus [EM84] zur stärkeren Störsignaldämpfung ohne Verlust an Sprachsignalqualität führt.

Das Verfahren aus [FG14] ist eine Weiterentwicklung des Verfahrens aus [GBM08] kombiniert mit einer statistischen Modellierung aus [BM10]. Dabei wird angenommen, dass spektrale Amplituden einer steilgipfligen Chi-Verteilung mit einem Formparameter unterliegen, der mit sich einen zusätzlichen Freiheitsgrad in die SPP-Schätzung mitbringt. Während in [BM10] eine SPP-Schätzung für einen einzigen Zeit-Frequenz-Punkt hergeleitet wird, stellt [FG14] die SPP-Berechnung vor, die wie in [GBM08] auf einer *a posteriori* SNR-Mittelung über die benachbarten Zeit-Frequenz-Punkte basiert. Das *a priori* SNR wird hier ähnlich wie in [GBM08] auf einen festen Wert gesetzt, der sowohl für eine lokale als auch eine globale Umgebung optimiert wird. In Experimenten mit den SPP-Schätzern aus [EM84, GBM08, BM10] stellt sich heraus, dass das vorgeschlagene Verfahren eine höhere Störsignaldämpfung erzielt ohne Verlust in der Sprachsignalqualität aufzuweisen und dabei die kleinste Menge von *musical tones* im prozessierten Signal produziert.

SPP-Schätzung mit weiteren Ansätzen: In [TVHU13] wird ein ausgeklügelter SPP-Schätzer basierend auf einem zweidimensionalen (2-D) gerichteten HMM hergeleitet, mit welchem die Zeit-Frequenz-Korrelationen von Sprachsignalen modelliert werden. Für jeden Zeit-Frequenz-Punkt eines Spektrogramms wird dabei ein verborgener binärer Zustand modelliert, welcher die *a posteriori* SNR-Werte als Beobachtungen emittiert, die wie in [GBM08] einer skalierten Chi-Quadrat-Verteilung unterliegen. Die benachbarten verborgenen Zustände sind miteinander mit Hilfe eines gerichteten Wahrscheinlichkeitsgraphen über Übergangswahrscheinlichkeiten verbunden. Somit besitzt das hier vorgestellte 2-D HMM eine viel stärkere Modellierungskraft als das eindimensionale HMM aus [SKS99]. Die Sprachpräsenzwahrscheinlichkeit ist dabei als *a posteriori* Wahrscheinlichkeit eines verborgenen Zustandes im 2-D HMM gegeben alle Beobachtungen eines Spektrogramms definiert. Um eine aufwändige Berechnung einer exakten Inferenz für so ein großes HMM zu vermeiden, wird auf das Turbo-Prinzip zurückgegriffen, welcher für Inferenzberechnung eine iterative Prozedur zwischen der horizontalen (Zeit) und vertikalen (Frequenz) Richtung eines 2-D HMM mit einem Austausch von sogenannten extrinsischen Informationen in einem Zwischenschritt vorsieht [Gla15]. Ein Vergleich der resultierenden *receiver operating characteristic* (ROC) Kurven zeigte, dass das vorgeschlagene Verfahren bereits nach einigen wenigen Turbo-Iterationen die Leistungsfähigkeit des SPP-Schätzers aus [GBM08] übertraf.

Laut Untersuchungen in [Gla15] konnte der SPP-Schätzer aus [TVHU13] eingesetzt im System zur einkanaligen Sprachsignalentstörung jedoch nicht überzeugen, was vermutlich daran lag, dass seine frequenzabhängigen Modellparameter hier mit einer einfachen Methode einer energiebasierten Zählung berechnet wurden. Allerdings kamen solche Modellparameter einem einfacheren SPP-Schätzer zugute, der die statistische Modellierung aus [GBM08] verwendet und dabei explizit keine ZF-Korrelationen der Sprachsignale berücksichtigt. D. h., er betrachtet die Beobachtungen der einzelnen ZF-Punkte als statistisch unabhängig, wie dies auch in einigen anderen Verfahren zur SPP-Schätzung in [MM80, EM84, MCA99, SKY99, DA09, FF12] der Fall ist. Die frequenzabhängigen *a priori* SAP $q(k)$ kalkulieren dabei eine tiefpasslastige Natur eines Sprachsignals ein und tragen dadurch zu einer guten SPP-Schätzung bei, die verglichen mit einem SPP-Schätzer konkurrenzfähig bleibt, der ein 2-D HMM mit den ungerichteten Wahrscheinlichkeitsgraphen verwendet.

In [MA$^+$16] wird ein Verfahren vorgestellt, das bei seiner SPP-Schätzung im aktuellen Zeit-Frequenz-Punkt nur eine begrenzte lokale Zeit-Frequenz-Umgebung berücksichtigt, die ähnlich wie in [GBM08] definiert wird. STFT-Koeffizienten dieser ZF-Umgebung werden hier zu einem Vektor mit den komplexwertigen Komponenten zusammengefasst, der mit den komplexwertigen multivariaten Normalverteilungen modelliert wird. Unter den Annahmen einer Kovarianzmatrix des ungestörten Sprachsignals von Rang 1 und einer stationären Stö-

Quelle	fixierte *a priori* SAP	geschätzte *a priori* SAP	Multiplikation mit G_{H_1}	Potenzierung von G_{H_1}	binäre VAD	weiche SPP	lokale ZF-Umgebung	globale ZF-Umgebung	SPP pro STFT-Rahmen	geschätzte *a priori* SNR	fixierte *a priori* SNR	Verwendung von HMM
1. [MM80]	✓		✓			✓				✓		
2. [EM84]	✓		✓			✓				✓		
3. [Yan93]	✓		✓		✓				✓	✓		
4. [SKS99]	✓				✓				✓	✓		✓
5. [KC00]	✓		✓			✓			✓	✓		
6. [MCA99]		✓	✓			✓				✓		
7. [SKY99]		✓	✓		✓	✓				✓		
8. [Coh01]		✓		✓		✓	✓	✓	✓	✓		
9. [SA05]			✓		✓		✓	✓				
10. [GBM08]	✓		✓			✓	✓	✓			✓	
11. [DA09]		✓	✓			✓	✓	✓	✓	✓		
12. [FF12]	✓		✓			✓				✓		
13. [TVHU13]	✓					✓					✓	✓
14. [FG14]	✓		✓			✓	✓	✓			✓	
15. [Gla15]	✓					✓					✓	
16. [MA$^+$16]	✓		✓			✓	✓			✓		

Tabelle 3.4.: Merkmale einiger Verfahren zur SPP-Schätzung.

rung wird hier ein recheneffizienter SPP-Schätzer hergeleitet. Im experimentellen Vergleich mit den Verfahren aus [Coh01] und [GBM08] erreicht der vorgeschlagene SPP-Schätzer bessere Leistungsfähigkeit, wenn die Kovarianzmatrix des Rauschens aus dem gegebenen Störsignal berechnet wird. Untersuchungen der Leistungsfähigkeit dieses Verfahrens für den Fall, dass die Kovarianzmatrix des Störsignals geschätzt wird, stehen noch aus.

Die charakteristischen Merkmale der in diesem Kapitel beschriebenen SPP-Schätzer sind in Tab. 3.4 zusammengefasst. Dabei fällt auf, dass viele SPP-Schätzer multiplikativ in die Berechnung der finalen Filterfunktion eingehen und eine weiche SPP-Schätzung realisieren. Nicht unerwähnt dürfen an dieser Stelle tiefe neuronale Netze bleiben, die besonders zuletzt eine hervorragende Leistungsfähigkeit in einer spektralen SPP-Schätzung zeigen. Einige Beispiele solcher DNN-basierter Maskenschätzer werden in Abschnitt 6.1 ausführlich diskutiert. Allerdings können neuronale Netze für eine spektrale Sprachsignalentstörung auch direkt eingesetzt werden und somit das System aus Abb. 2.2 komplett ersetzen.

3.5. Spektrale Sprachsignalentstörung mit künstlichen neuronalen Netzen

Dank den jüngsten Fortschritten in der Forschung erlebt die Mustererkennung unter Verwendung von tiefen neuronalen Netzen eine wahre Renaissance [LBH15]. Bereits in [HSW89] wird postuliert, dass tiefe *feedforward* Netze mit einer ausreichenden Neuronenanzahl beliebige funktionale Zusammenhänge zwischen ihrer Eingangsschicht und ihrer Ausgangsschicht mit einer beliebigen Genauigkeit approximieren und somit als universelle Approximatoren gelten können. Einen großen Beitrag zu Verwirklichung dieser Behauptung lieferte die Formulierung eines weit verbreiteten Verfahrens der Fehlerrückführung (engl. *backpropagation*) zum Training der DNN-Parameter unter Verwendung des stochastischen Gradientenabstiegs (engl. *stochastic gradient descent*) [RHW86]. Um zeitliche Korrelationen der beteiligten Zufallsprozesse, die in einem Sprachsignal zwangsläufig vorhanden sind, in einem neuronalen Netz abbilden zu können, wird in [Pin87] ein rekurrentes neuronales Netz vorgestellt, dessen spezielle Variante die *long-short-term memory* (LSTM) Netze sind [HS97]. Als hinreichende Voraussetzungen für die Realisierung von guten Approximationseigenschaften der Netze in einer Anwendung werden erstens die Verwendung von leistungsfähigen schnellen Rechnern, zweitens das Vorhandensein von großen Mengen an Trainigsdaten und drittens eine sinnvolle Initialisierung der Gewichte eines Netzes gefordert [HS06]. Das Miteinbeziehen von *graphics processing units* (GPU) beschleunigte das sonst sehr zeitaufwendige DNN-Training durch die massive Parallelisierung der notwendigen Rechenoperationen und erweiterte den Einsatz von Netzen auf die sogenannten *large-scale applications* [RMN09]. Ein guter Rückblick auf den Werdegang von künstlichen neuronalen Netzen mit einem guten Überblick über ihre Tiefen und Höhen ist in [LBH15] zu finden.

Frühere Systeme unter Verwendung von neuronalen Netzen: In den früheren Veröffentlichungen zur DNN-basierten einkanaligen spektralen Sprachsignalentstörung werden noch sehr kleine und einfache *feedforward* Netze eingesetzt. So wird in [XVC93] eine DNN-basierte spektrale Filterfunktion geschätzt, die für eine Merkmalsentstörung verwendet wird. In [PG04] wird im Unterschied zu [XVC93] ein DNN für eine direkte Schätzung der spektralen Merkmale eines ungestörten Sprachsignals im Rahmen einer sogenannten Regression

eingesetzt. Unter dem Begriff einer Regression wird verstanden, dass die spektralen Merkmale eines ungestörten Sprachsignals vom neuronalen Netz direkt aus den des gestörten Sprachsignals berechnet werden. Dabei wird hier ein rekurrentes neuronales Netz (RNN) zur spektralen Sprachsignalentstörung verwendet. Bedingt durch die Rückkopplungen, die in der Architektur des RNN vorhanden sind, ist ein RNN im Stande, zeitliche Korrelationen der Sprachsignale zu erfassen. Außerdem wird das RNN hier mit einer einfachen energiebasierten VAD kombiniert, die in den experimentellen Ergebnissen die Leistungsfähigkeit des Netzes steigert, besonders hinsichtlich Rauschunterdrückung bei positiven Werten des eingangsseitigen globalen $\mathrm{SNR}_{\mathrm{IN}}$. In den Experimenten übertreffen sowohl die DNN-basierte Filterfunktion als auch das RNN kombiniert mit einer VAD eine konventionelle spektrale Subtraktion wie in [BSM79]. Eine gute Übersicht über die früheren Systeme zur Sprachsignalentstörung unter Verwendung von neuronalen Netzen ist in [WN99] zu finden.

A priori SNR-Schätzer: Auch das *a priori* SNR wird mit Hilfe von Netzen geschätzt. In [SLF11] werden zwei getrennte *feedforward* Netze für zwei Szenarien trainiert, das eine für Sprachsignalanwesenheit und das anderer für Sprachsignalabwesenheit. Dabei werden an die zwei Eingänge der Netze die beiden Komponenten des DD-Verfahrens aus (3.12) und (3.14), gemessen in Dezibel, angelegt, aus denen das *a priori* SNR, gemessen auch in Dezibel, für das jeweilige Szenario berechnet wird. Somit hat jedes Netz nur einen einzigen Knoten in der Ausgangsschicht, wie dies auch in [XVC93] der Fall ist. Sonst beinhalten die Netze drei verborgene Schichten mit nur jeweils 10 Neuronen in jeder Schicht. Man beachte, dass die *a priori* SNR-Schätzwerte der beiden Netze mit Hilfe eines SPP-Schätzers zu einem finalen *a priori* SNR-Schätzwert kombiniert werden, der im Rahmen der experimentellen Untersuchungen in drei konventionellen spektralen Filterfunktionen verwendet wird. Für die Schätzung des Rauschleistungsdichtespektrums wird hier das MS-Verfahren aus [Mar01] verwendet. Somit wird das neuronale Netz unterstützend in einer modellbasierten Sprachsignalentstörung eingesetzt und liefert eine robuste *a priori* SNR-Schätzung, die als ein dominanter Parameter der spektralen Filterfunktionen betrachtet wird [Cap94]. In den Experimenten führte der DNN-basierte *a priori* SNR-Schätzer zu den entstörten Sprachsignalen mit sowohl besserer Qualität als auch höherer Rauschunterdrückung.

Denoising Autoencoder: In [LTMH13] wird eine Entstörung mittels einer Regression von Mel-spektralen Koeffizienten (MSK) unter Verwendung eines *denoising* Autoencoders realisiert, der aus drei verborgenen *feedforward* Schichten mit jeweils 500 Knoten besteht. Während ein Autoencoder in erster Linie für eine effiziente Kodierung verwendet werden kann, sind die *denoising* Autoencoder im Stande, Regressionsaufgaben erfolgreich zu lösen. Bei gleichzeitiger Verwendung von Mel-Filterkoeffizienten und einer langen Eingangsschicht mit 440 Eingangsknoten gelingt es hier, einen zeitlichen Kontext von 11 aufeinander folgenden Rahmen in die Schätzung miteinzubeziehen. Die Anzahl der DNN-Parameter ist hier relativ groß im Vergleich zu den gerade erwähnten Netzen, was mit dem jüngsten Durchbruch in der Sprachsignalverarbeitung mit neuronalen Netzen einhergeht. Aus diesem Grund werden für ein erfolgreiches Training des Netzes auch mehr Daten benötigt. Beim Training des Autoencoders mit der Technik aus [HOT06] werden auf die Eingangsschicht die MSK des verrauschten Sprachsignals und auf die Ausgangsschicht die des ungestörten Sprachsignals angelegt. Da die gestörten Sprachsignale in der Testphase mit den Störsignalen der selben Rauschtypen überlagert werden, die im Training des Autoencoders verwendet werden, gibt es hier keine Diskrepanz zwischen der Trainigs- und Testphasen. Bemerkenswert ist, dass die experimentelle Untersuchungen in [LTMH13] zeigten, dass der *denoising* Autoencoder ein

konventionelles System zur spektralen Entstörung aus [Coh03] hinsichtlich der Qualität der entstörten Sprachsignale bei weitem übertrifft, jedoch hinsichtlich der Rauschunterdrückung in etwa die gleiche Leistungsfähigkeit aufweist.

DNN-basierte Regression: Ermutigt von der Möglichkeit tiefere und größere Netze trainieren zu können, werden weitere DNN-basierte Systeme entwickelt wie in [XDDL14], die auch dann eine gute Leistungsfähigkeit aufweisen, wenn sie Sprachsignale entstören, die mit den Rauschtypen überlagert werden, die das Netz in der Trainigsphase nicht verwendet wurden. Dabei findet hier die Sprachsignalentstörung wieder im Bereich von normalisierten logarithmischen Leistungsdichtespektren ähnlich wie in [XVC93] statt, wofür ein *feedforward* Netz mit drei verborgenen Schichten mit jeweils 2^{11} Knoten eingesetzt wird. Das Netz wird mit den Sprachsignalen im Umfang von 100 Stunden mit dem Verfahren aus [HOT06] trainiert und liefert die entstörten Signale mit besserer Sprachsignalqualität gemessen in PESQ-Werten als das konventionelle OMLSA-Verfahren aus [CB01b]. Allerdings wird dabei keinen Wert auf eine kausale Signalverarbeitung gelegt, was dem DNN-basierten System einen unfairen Vorteil gegenüber dem konventionellen System verschafft. Bemerkenswert ist dennoch, dass das DNN in [XDDL14] wie auch der Autoencoder in [LTMH13] sich als eigenständige Systeme zur spektralen Entstörung darstellen und ganz ohne modellbasierte Schätzer aus Abb. 2.2 auskommen. Somit stellt sich eine interessante Frage, ob die konventionellen Systeme in ihrer Existenz bedroht sind. Bei der Suche nach einer Antwort liefert [MT16] einen bemerkenswerten Beitrag, in dem ein DNN sowohl als eine Regression als auch in Kombination mit den konventionellen Verfahren eingesetzt wird.

Vergleich verschiedener Architekturen: In [MT16] werden verschiedene DNN-basierte Systeme zur einkanaligen spektralen Sprachsignalentstörung mit einem konventionellen modellbasierten Verfahren verglichen. Dabei werden die Schwerpunkte der Untersuchung auf eine kausale Signalverarbeitung und auf Filterung von Sprachsignalen gestört von den Rauschtypen, die dem Netz in der Trainigsphase nicht präsentiert werden, gelegt. In einem der Systeme übernimmt ein DNN die komplette Aufgabe der spektralen Entstörung mittels einer Regression im STFT-Bereich und berechnet direkt einen Schätzwert der spektralen Amplituden des ungestörten Sprachsignals $|\hat{S}(k,\ell)|$ aus den gestörten spektralen Amplituden[11]. Dieser Ansatz führt zwar zu den prozessierten Signalen mit bester Qualität allerdings nur bei einer antikausalen Entstörung von Sprachsignalen, die von den Rauschtypen gestört werden, welche dem DNN in der Trainingsphase präsentiert werden. Bei einer kausalen Signalverarbeitung von Sprachsignalen gestört von anderen Rauschtypen, verschlechtert sich die Signalqualität deutlich und fällt sogar unter das Niveau der konventionellen spektralen Entstörung, die mit den modellbasierten Verfahren wie in Abb. 2.2 aufgebaut wird. Die Leistungsfähigkeit der DNN-basierten kausalen Entstörung verbessert sich, wenn das DNN statt der spektralen Amplituden $|\hat{S}(k,\ell)|$ die Filterfunktion $G(k,\ell)$ aus (2.17) lernt, mit deren Hilfe die STFT-Koeffizienten des gestörten Sprachsignals multiplikativ entstört werden. Allerdings übertrifft dieses System die konventionelle spektrale Entstörung hinsichtlich der Sprachsignalqualität gemessen in PESQ-Werten nur leicht.

Eine weitere signifikante Steigerung der Leistungsfähigkeit der DNN-basierten Entstörung wird in [MT16] dadurch erreicht, dass ein Hybridsystem aufgebaut wird, in dem drei verschiedene neuronale Netze zum Einsatz kommen. Während das erste DNN eine modell-

[11] Man beachte, dass die Eingangsdaten für eine Verarbeitung in einem neuronalen Netz in der Regel normalisiert werden, sodass sie mittelwertfrei werden und die Varianz gleich 1 aufweisen [XDDL14]. Ausgangsseitig wird diese Normalisierung rückgängig gemacht.

basierte RLDS-Schätzung unterstützt, übernehmen die beiden anderen neuronalen Netze die Aufgaben der Bausteine 2 und 3 wie im System in Abb. 2.2. Die RLDS-Schätzung $\hat{\lambda}_D(k,\ell)$ wird dabei mittels einer rekursiven Glättung wie in (2.48) mit einem zeitvarianten Glättungsparameter realisiert, der vom DNN-basierten VAD-Schätzer gesteuert wird. Das zweite DNN schätzt das Leistungsdichtespektrum des ungestörten Sprachsignals $\hat{\lambda}_S(k,\ell)$ mittels Regression. Aus $\hat{\lambda}_D(k,\ell)$ und $\hat{\lambda}_S(k,\ell)$ werden anschließend die Schätzwerte für das *a posteriori* SNR $\hat{\gamma}(k,\ell)$ und für das *a priori* SNR $\hat{\xi}(k,\ell)$ berechnet, die als Eingangsgrößen des dritten neuronalen Netzes dienen, das die Filterfunktion $G(k,\ell)$ bestimmt. Somit wird ähnlich wie bei einer modellbasierten Entstörung die Filterfunkton $G(k,\ell)$ aus $\hat{\xi}(k,\ell)$ und $\hat{\gamma}(k,\ell)$ berechnet. Dabei wird der resultierende funktionale Zusammenhang $G = f(\xi,\gamma)$ nicht durch die statistische Modellierung festgelegt, sondern datengetrieben gelernt ähnlich wie in [XVC93]. Im Unterschied zu den ersten beiden Netzen, die jeweils drei verborgene *feed-forward* Schichten mit 2^{11} Neuronen beinhalten, kommt dieses dritte Netz mit nur einer einzigen verborgenen Schicht bestehend aus 2^{10} Neuronen aus. Während der modulare Aufbau des DNN-basierten Hybridsystems die Reduzierung der Komplexität der einzelnen neuronalen Netze ermöglicht, trägt die Verwendung des modellbasierten robusten RLDS-Schätzers zur besseren Generalisierbarkeit der DNN-basierten kausalen Sprachsignalentstörung hinsichtlich der für das Netz unbekannten Rauschtypen bei. Als Fazit lässt sich festhalten, dass die modellbasierten Verfahren und die DNN-basierten Systeme einander sehr gut ergänzen. Während die DNN-basierte Signalverarbeitung zu den robusten Systemkomponenten führt, sorgen modellbasierten Verfahren für eine gute Generalisierbarkeit der Systeme.

Weitere Anwendungen von neuronalen Netzen: Die neuronalen Netze werden allerdings nicht nur zur Lösung von Regressionsaufgaben, als *a priori* SNR-Schätzer oder zur Approximation von datengetriebenen Filterfunktionen eingesetzt. Eine weitere prominente DNN-Anwendung in der spektralen Sprachsignalverarbeitung ist die Schätzung der Sprachpräsenzwahrscheinlichkeit, auf die in Kap. 6 eingegangen wird.

4. Wissenschaftliche Ziele

Digitale Verarbeitung von gestörten Sprachsignalen kann unterschiedliche Ziele verfolgen:

1. Steigerung der Verständlichkeit eines Sprachsignals,

2. Verbesserung der akustischen Qualität eines Sprachsignals,

3. Reduzierung des Störsignalanteils oder Rauschunterdrückung.

In diesem Kapitel wird erläutert, welche Beiträge diese Arbeit zum Erreichen der oben erwähnten Zielen leistet.

Anfangs glaubte man, dass die konventionellen Systeme zur Sprachsignalentstörung die Verständlichkeit des Sprachsignals nicht verbessern können, und konzentrierte sich aus diesem Grund auf die Entwicklung von Systemen, welche die Ziele 2 und 3 verfolgten [LO79, HL07]. Dabei erkannte man, dass bei solchen Systemen die Verbesserung der Sprachsignalqualität und die Reduzierung des Störsignalanteils zwei konkurrierenden Ziele sind [MM80, CG08]. Diese Erkenntnis macht die Entwicklung der einzelnen Bausteine eines Systems zur Entstörung spektraler Amplituden wie in Abb. 2.2 oft zur Suche nach einem besseren Kompromiss zwischen einer bestmöglichen Sprachsignalqualität und einer möglichst guten Rauschunterdrückung, wobei das Hauptaugenmerk häufig auf hoher Qualität der entstörten Sprachsignale liegt. Wird bei der Entwicklung eines Verfahrens im Vergleich zu den bereits existierenden Verfahren höhere Störsignaldämpfung ohne Verluste in der Sprachsignalqualität beobachtet, kann man vom einem qualitativ neuen Verfahren sprechen [YKR05, EMTF15]. Führt ein Verfahren sowohl zur besseren Sprachsignalqualität als auch zur stärkere Störsignaldämpfung (Ziele 2 und 3 gleichzeitig erreicht), kann man von einem besonderen Fortschritt sprechen.

Alle vier in Abschnitt 2.2 beschriebenen Bausteine spektraler Entstörung können zum Erreichen der oben erwähnten Ziele in verschiedener Art und Weise beitragen. Die RLDS-Schätzung hat dabei eine besondere Stellung direkt am Eingang des Systems in Abb. 2.2 und beeinflusst somit zwangsläufig die Leistungsfähigkeit aller weiteren Systembausteine. Diese Tatsache dient für diese Arbeit als Motivation, sich mit der Entwicklung der Verfahren zur RLDS-Schätzung ausgiebiger auseinander zu setzen. Somit setzt sich diese Arbeit neben dem einleitenden Teil A aus zwei weiteren Hauptteilen zusammen. Während in Teil B drei RLDS-Schätzer entwickelt werden, die unterschiedliche oben erwähnten Ziele verfolgen, werden in Teil C drei Verfahren vorgestellt, die jeweils in den Bausteinen 2, 3 und 4 eingesetzt werden und alle zum Erreichen des zweiten und des dritten Ziels beitragen.

4.1. Schätzung spektraler Rauschleistungsdichte

Durch die besondere Stellung im System zur Entstörung spektraler Amplituden in Abb. 2.2 darf die RLDS-Schätzung als Baustein 1 eine erhöhte Aufmerksamkeit in dieser Arbeit genießen. So werden in Teil B der Arbeit drei verschiedene RLDS-Schätzer mit unterschiedlichen Zielsetzungen vorgestellt und zwar:

- ein modifiziertes *Minimum Statistics* Verfahren mit dem Ziel 1,
- ein vom neuronalen Netz unterstützter Schätzer mit den Zielen 2 und 3,
- ein Bayesscher Postprozessor mit dem Ziel 2.

Alternative Steuerungsfunktion für das *Minimum Statistics* Verfahren

Für viele Anwendungen, wie z. B. die automatische Spracherkennung, ist die Verständlichkeit des prozessierten Sprachsignals von weit größerer Bedeutung als die Signalqualität oder die Menge des Restrauschens. Im Unterschied zu den früheren Untersuchungen gelang es erst vor einigen Jahren mit der Entwicklung von binären spektralen Filterfunktionen wie in [KLHL09], die Verständlichkeit des prozessierten Signals zu verbessern. Dabei hat man festgestellt, dass die Minimierung des mittleren quadratischen Fehlers, die bei der Entwicklung von vielen Schätzern oft angestrebt wird, ein ungeeignetes Optimierungskriterium hinsichtlich der Steigerung der Verständlichkeit des Sprachsignals ist [LK11]. Die Verwendung einer binären Filterfunktion führt zu einer 'harten' Entscheidung, ob ein Zeit-Frequenz-Punkt unverarbeitet das System passieren kann oder seine spektrale Amplitude zu Null gesetzt wird[1]. Somit wird auf die Transitionen in der Filterfunktion verzichtet, wo sie die reellen Werte im Bereich zwischen 0 und 1 einnimmt.

Basierend auf dieser Beobachtung stellt sich die Frage, mit welchen Mitteln Transitionen der konventionellen spektrale Filterfunktionen reduziert werden können, sodass die Filterfunktionen zwischen G_{min} und G_{max} schneller schalten. Wie in Abschnitt 2.2 erwähnt und in Abb. 2.2 dargestellt, können spektrale Filterfunktionen der konventionellen Sprachsignalentstörung von der *a posteriori* SNR $\gamma(k,\ell)$, von der *a priori* SNR $\xi(k,\ell)$ und von der SPP $\mathcal{P}(k,\ell)$ abhängig sein. Allerdings beinhalten die Schätzer von $\xi(k,\ell)$ und $\mathcal{P}(k,\ell)$ üblicherweise eine Glättung oder eine Art von Mittelung, um verlässliche Schätzwerte zu erreichen. Als ein wirksames Mittel, der Filterfunktion die gewünschten Eigenschaften einzuprägen, bleibt somit nur $\gamma(k,\ell)$, die ja aus einem RLDS-Schätzwert berechnet wird. Somit wird ein dynamischer und gleichzeitig robuster RLDS-Schätzer gesucht.

Wie die Analyse der modernen RLDS-Verfahren zeigte, verzichtet nur das *Minimum Statistics* Verfahren auf eine ausgangsseitige Glättung und gibt somit die Möglichkeit, mehr Dynamik in die RLDS-Schätzung zu bringen, ohne dabei auf eine robuste Schätzung zu verzichten. Eine Möglichkeit, das MS-Verfahren zu modifizieren, wird in Kap. 5 vorgestellt. Hier wird eine alternative Kontrollfunktion entwickelt, welche zur Verbesserung der Verständlichkeit entstörter Sprachsignale gemessen mit Hilfe des STOI-Maßes führen soll.

[1] Durch das Setzen der spektralen Amplituden zu Null und somit dem Verzichten auf die Verwendung von G_{min} wird im prozessierten Sprachsignal eine große Menge des musikalischen Rauschens produziert. Allerdings leidet darunter die Sprachverständlichkeit nicht wie die Untersuchungen in [LK11] zeigten.

RLDS-Schätzung unter Verwendung eines neuronalen Netzes

Wie in Kap. 1 und in Abschnitt 3.5 erwähnt, genießt seit einigen Jahren die Verwendung von tiefen neuronalen Netzen in der digitalen Sprachsignalverarbeitung besondere Aufmerksamkeit der Forscher [HCS+07]. Als Folge werden auch für die Verarbeitung von gestörten Sprachsignalen DNN-basierte Systeme entwickelt, welche zwar eine ausgiebige Trainingsphase mit vielen Daten benötigen, dafür aber von der Leistungsfähigkeit her die Systeme mit den modellbasierten Verfahren oft übertreffen. Dabei können die Netze die Rolle der nichtlinearen Filter übernehmen, welche eine Abbildung von den gestörten Sprachsignalen auf die ungestörten Signale realisieren [XDDL14]. Oder sie werden mit den modellbasierten Verfahren in den Hybridsystemen erfolgreich kombiniert [MT16]. Auch am Fachgebiet Nachrichtentechnik der Universität Paderborn werden solche Systeme momentan entwickelt, die unter anderem für die Entstörung von mehrkanaligen Sprachsignalaufnahmen verwendet werden wie in [HDCHU15]. Hier übernimmt das neuronale Netz die Aufgabe, die Zeit-Frequenz-Punkte zu finden, die entweder vom Sprachsignal oder vom Störsignal dominiert werden.

Bei näherer Betrachtung von modernen RLDS-Schätzern in Abschnitt 3.1 wurde festgestellt, dass die VAD/SPP Schätzung als eine bewährte modellbasierte Technik in der RLDS-Schätzung häufig eingesetzt wird. Somit liegt der Gedanke nahe, das neuronale Netz aus [HDCHU15] für diese Technik einzusetzen und mit einer ausgangsseitigen Glättung zu kombinieren. Dabei soll das neuronale Netz nur die Zeit-Frequenz-Punkte finden, in denen das Störsignal allein die spektrale Signalamplitude dominiert, indem es die Wahrscheinlichkeit für die alleinige Präsenz des Störsignals (engl. *noise-only presence probability*, NPP) berechnet. Dafür wird das neuronale Netz aus [HDCHU15] modifiziert für eine kausale Signalverarbeitung verwendet. Der resultierende DNN-basierte RLDS-Schätzer wird in Kap. 6 vorgestellt und in einer experimentellen Untersuchung mit den modernen Verfahren für RLDS-Schätzung verglichen, die für den fairen Vergleich auf den Daten der Datenbank optimiert werden, auf dem das verwendete neuronale Netz trainiert wird.

Bayesscher Postprozessor zur RLDS-Schätzung

Im Gegensatz zu den ASR-Anwendungen, in denen die Verständlichkeit der prozessierten Sprachsignale im Vordergrund steht, spielt bei der Sprachsignalübertragung die Qualität der prozessierten Signale eine wichtige Rolle, wofür entsprechende RLDS-Schätzer gesucht werden. Dies ist die Motivation für die Entwicklung eines Postprozessors, der dazu dienen soll, die RLDS-Schätzwerte bestehender Rauschschätzer so zu verbessern, dass die Qualität der prozessierten Signale steigt. Dafür soll zunächst ein radikaler Paradigmenwechsel hinsichtlich der Definition vollzogen werden, welcher Anteil im gestörten Sprachsignal $y(n) = s(n) + d(n)$ unerwünscht ist und welcher nicht.

In der Regel wird das Störsignal $d(n)$ als eine unwillkommene Signalkomponente betrachtet, welche im Rahmen der spektralen Filterung entfernt werden soll. Allerdings stellt sich für einen RLDS-Schätzer das ungestörte Sprachsignal $s(n)$ als ein Hindernis für eine adäquate RLDS-Schätzung dar, denn wegen $s(n)$ kann $d(n)$ nicht direkt beobachtet werden. Mit dieser Sichtweise lässt sich ein Bayesscher RLDS-Schätzer herleiten, dessen *a priori* Wissen im aktuellen Signalblock auf dem RLDS-Schätzwert des vorigen Blocks beruht und die Beobachtungsverteilung auf dem LDS des ungestörten Sprachsignals basiert, das als

gegeben vorausgesetzt wird. In einer konkreten Anwendung muss also das LDS des ungestörten Sprachsignals in einer Vorstufe geschätzt werden, zu welcher der zu entwickelnde Bayesscher RLDS-Schätzer als Postprozessor agiert.

Realisiert werden kann dies mit Hilfe eines vorgeschalteten *a priori* SNR-Schätzers, der selbst eine eigene RLDS-Schätzung benötigt, für welche ein beliebiger bereits bekannter RLDS-Schätzer verwendet werden kann. Somit wird auch der Unterschied des vorgeschlagenen Postprozessors zu den in Abschnitt 3.1 vorgestellten Bayeschen RLDS-Schätzern klar [Yu09, HHJ10], die als eigenständige Schätzer entwickelt wurden und somit auch in der Vorstufe eingesetzt werden können. Bemerkenswert ist, dass der vorgeschlagene Paradigmenwechsel zu einer Bayes-motivierten ausgangsseitigen Glättung führt, die sich für die Qualität der entstörten Signale als besonders günstig erweist, wie in Kap. 7 gezeigt wird.

4.2. Generalisierte modellbasierte Sprachsignalentstörung

Im Teil C dieser Arbeit werden drei modellbasierte Verfahren (jeweils eins für die Bausteine 2, 3 und 4) entwickelt, die im SNR-Bereich eines Systems zur Entstörung spektraler Amplituden angesiedelt sind, wie Abb. 2.2 zeigt, und zwar:

- Baustein 2: MAP-Schätzer generalisierter log-spektraler Amplituden
- Baustein 3: Generalisiertes *Decision-Directed* Verfahren
- Baustein 4: SPP-Schätzung im generalisierten SNR-Bereich.

Dabei wird die Verbesserung der akustischen Qualität des prozessierten Sprachsignals in den Vordergrund gestellt und somit das Ziel 2 verfolgt.

Im Unterschied zu den konventionellen Verfahren werden bei diesen Verfahren z. B. statt spektraler Amplituden $|S(k,\ell)|$ oder $|N(k,\ell)|$ die generalisierten spektralen Amplituden $|S(k,\ell)|^\beta$ oder $|N(k,\ell)|^\beta$ mit einem Kompressionsfaktor $\beta \in \mathbb{R}_{>0}$ für die statistische Modellierung verwendet [STCT98]. Zwei Leitgedanken motivieren dabei die Einführung von GSA in dieser Arbeit. Der erste ist in einer Aussage von George E. P. Box verankert, dass alle statistischen Modelle im wesentlichen falsch sind [Box79][2]. Der Grund für das Versagen der Modelle sind Annahmen und Vereinfachungen, die bei der Modellierung der Daten getroffen werden, in der Realität jedoch nicht (oder nicht ganz) stimmen. So wird in der Sprachsignalverarbeitung häufig angenommen, dass die spektralen Amplituden der benachbarten Zeit-Frequenz-Punkte eines Sprachsignals unkorreliert oder sogar statistisch unabhängig sind. Dies ist allerdings dank der Signalverarbeitung mit überlappenden Fenstern selbst beim weißen Rauschen falsch, geschweige denn beim Sprachsignal.

Ein möglicher Weg, die fehlerhaften Annahmen zu berücksichtigen, ohne dabei die Modelle sehr kompliziert zu machen, ist die Modellierung von spektralen Amplituden mit Modellen, die für generalisierte spektrale Amplituden gelten. Auf diese Weise wird höhere Flexibilität der statistischen Modelle erreicht, die durch β einen zusätzlichen Freiheitsgrad bekommen, um die zu verarbeitenden Daten statistisch besser zu beschreiben. In einer Optimierungsphase wird dann β so gewählt, dass die Daten mit den Modellen bzgl. eines festgelegten Kriteriums am besten beschrieben werden. Im gewissen Sinne werden die falsch

[2] In Englisch vermittelte George E. P. Box in [Box79]: 'Essentially, all models are wrong, but some are useful'.

getroffenen Annahmen dabei zum Teil kompensiert, sodass Modelle etwas 'richtiger' (oder laut Box 'nützlicher') werden als zuvor. Der zweite Leitgedanke beruht auf einer Hypothese, dass es keine Garantie gibt, dass der Bereich, in dem die Größen definiert werden, auch der bestmögliche Bereich für ihre Schätzung ist.

MAP-Schätzer generalisierter log-spektraler Amplituden

Diesen Leitgedanken folgend, wird im Rahmen dieser Arbeit eine neue spektrale Filterfunktion entwickelt, die laut den experimentellen Untersuchungen im Vergleich zur leistungsfähigen und weit verbreiteten MMSE-LSA-Funktion zur besseren Rauschunterdrückung führt, ohne Verluste in der Sprachsignalqualität aufzuweisen. Verglichen mit der MMSE-GSS-Filterfunktion erreicht sie bessere Signalqualität mit einer leichten Verbesserung der Störsignaldämpfung. Die vorgeschlagene Filterfunktion nutzt den Synergieeffekt, der bei Kombination des Bereichs der logarithmischen spektralen Amplituden $\ln|S(k,\ell)|$ mit dem Bereich der generalisierten spektralen Amplituden $|S(k,\ell)|^\beta$ entsteht. Diese Kombination ist sinnvoll aus folgenden Überlegungen:

- Die in Abschnitt 2.2 eingeführte MMSE-LSA-Filterfunktion, welche die Kostenfunktion $\mathbb{E}[(\ln \mathrm{SPP}|\hat{S}(k,\ell)| - \ln|S(k,\ell)|)^2]$ minimiert, stellt sich im Bezug auf die Qualität der entstörten Signale als eine der besten Funktionen dar [HL06]. Eine Analyse der MMSE-SA- und MMSE-LSA-Filterfunktionen in [EM85] schreibt die guten Eigenschaften des MMSE-LSA-Schätzers der Tatsache zu, dass er im Bereich der logarithmischen spektralen Amplituden definiert wird, der für die spektrale Sprachsignalverarbeitung hinsichtlich der subjektiven Wahrnehmung der Sprachsignalqualität vorteilhaft zu sein scheint [GBGM80].

- Die MMSE-GSS-Filterfunktion, welche einer parametrischen Formulierung des Schätzers entspringt und die Kostenfunktion $\mathbb{E}[(|\hat{S}(k,\ell)|^\beta - |S(k,\ell)|^\beta)^2]$ minimiert, erweist sich als konkurrenzfähig mit der MMSE-LSA-Filterfunktion bezüglich der Störsignaldämpfung [STCT98]. Bei einer nichtparametrischen Formulierung führt die Adaption des Kompressionsfaktors zur weiteren Verbesserung der Leistungsfähigkeit des MMSE-GSS-Schätzers allerdings auf Kosten der Verwendung spezieller mathematischer Funktionen wie der Bessel-Funktion oder der konfluenten hypergeometrischen Funktion [YKR05].

Die Kombination der beiden Bereiche ist dadurch möglich, dass spektrale Amplituden zunächst generalisiert und anschließend noch logarithmiert werden, was in $\ln|S(k,\ell)|^\beta$ resultiert. Auf diese Weise entstehen logarithmische generalisierte spektrale Amplituden (LGSA), die bei einer Bayesschen Schätzung verwendet werden können. Um einen Schätzer zu bekommen, der ohne spezielle mathematische Funktionen auskommt, die in den MMSE-Filterfunktionen vorkommen können, wird ein Maximum A-Posteriori basierter LGSA-Schätzer entwickelt. Dafür werden die Verteilungen der beteiligten Zufallsvariablen mit Hilfe der konsistenten Normalverteilungen approximiert, um zu einer recheneffizienten Lösung zu gelangen. Eine Alternative wäre einen MMSE-LGSA-Schätzer zu finden, welcher den *mean squared error* (MSE) $\mathbb{E}[(\ln|\hat{S}(k,\ell)|^\beta - \ln|S(k,\ell)|^\beta)^2]$ minimiert. Die MAP-LGSA-Filterfunktion wird in Kap. 8 hergeleitet und experimentell untersucht.

Generalisiertes *Decision-Directed* Verfahren

Auch bei der *a priori* SNR-Schätzung erweist sich die Verarbeitung im Bereich der generalisierten Größen als vorteilhaft. Da die *a priori* SNR laut (2.22) im SNR-Bereich definiert ist, wird sie häufig auch in diesem geschätzt. Ein prominentes Beispiel dafür ist das weit verbreitete *Decision-Directed* Verfahren, das in Abschnitt 3.3 ausführlich beschrieben wurde [EM85, Cap94]. Das DD-Verfahren berechnet das *a priori* SNR aus den Schätzwerten der *a posteriori* SNR und arbeitet somit ausschließlich im SNR-Bereich, der für eine *a priori* SNR-Schätzung nicht zwangsläufig optimal sein muss.

Um diese Hypothese genauer zu untersuchen, wird in Kap. 9 über die Einführung eines beliebigen Kompressionsfaktors der spektralen Amplituden ein Bereich der generalisierten (oder verallgemeinerten) SNR-Größen definiert. In diesem wird das generalisierte *a posteriori* SNR mit Hilfe der Weibull-Verteilung statistisch modelliert, in die das gesuchte *a priori* SNR als ein Parameter einfließt. Aus diesen Modellen gelingt es, ein generalisiertes DD-Verfahren herzuleiten, das im Vergleich zum konventionellen DD-Verfahren einen zusätzlichen Freiheitsgrad bekommt. Dieser kann in den experimentellen Untersuchungen mit Sprachsignalen, die vom weißen Rauschen gestört werden, hinsichtlich der Qualität der prozessierten Signale optimiert werden. Dabei stellt sich heraus, dass der konventionelle SNR-Bereich für das DD-Verfahren nur für die sehr stark gestörten Sprachsignale optimal ist, bei denen das globale SNR bei etwa $0\,\mathrm{dB}$ liegt oder noch kleiner ist. Für die Signale mit dem höheren globalen SNR wird das DD-Verfahren von seiner verallgemeinerten Version übertroffen. Das generalisierte DD-Verfahren wird ausführlich in Kap. 9 beschrieben, wo es in den umfangreichen Simulationen auch experimentell untersucht wird.

SPP-Schätzung im generalisierten SNR-Bereich

Wie in Abschnitt 3.4 bereits gezeigt, werden bei modernen Verfahren zur SPP-Schätzung die Korrelationen zwischen den STFT-Koeffizienten der benachbarten Zeit-Frequenz-Punkte eines Sprachsignals gewinnbringend ausgenutzt, wodurch eine robuste Schätzung erreicht wird [Coh01, GBM08, TVHU13, FG14, MA$^+$16]. Dabei findet die SPP-Schätzung häufig im SNR-Bereich statt wie in [Coh01, GBM08, FG14]. So wird in [GBM08] und [FG14] das *a posteriori* SNR als eine Eingangsgröße für SPP-Schätzung gewählt, die über eine bestimmte begrenzte Nachbarschaft des betrachteten Zeit-Frequenz-Punktes gleichgewichtet gemittelt wird. Allerdings nehmen die Korrelationen mit der Entfernung vom aktuellen Zeit-Frequenz-Punkt stark ab, wie einige Untersuchungen zeigten [Coh05, CG08, BCH11a]. Diese Tatsache bleibt bei den erwähnten konventionellen Verfahren leider unberücksichtigt, was zu einer fehlerbehafteten SPP-Schätzung führen kann. Um Beobachtungen einer ZF-Umgebung bei einer SPP-Schätzung unterschiedlich gewichten zu können, soll ein Verfahren entwickelt werden, das darauf speziell ausgelegt ist.

Dabei lässt sich das Konzept der generalisierten SNR-Größen mit einem zusätzlichen Freiheitsgrad auch in der SPP-Schätzung anwenden. Dabei ist nur die Frage zu klären, mit welcher Verteilung die gewichteten generalisierten *a posteriori* SNR-Größen modelliert werden sollen, wenn die einzelnen Summanden Weibull-verteilt sind. Eine Abhilfe soll eine Approximation mittels einer Normalverteilung sein, welche die angedachte SPP-Schätzung ermöglicht. Ausführliche Herleitung und experimentelle Untersuchung des neuartigen SPP-Schätzers im Bereich der generalisierten *a posteriori* SNR werden in Kap. 10 vorgestellt.

B. SCHÄTZUNG SPEKTRALER RAUSCHLEISTUNGSDICHTE

5. Alternative Steuerungsfunktion für das *Minimum Statistics* Verfahren

In diesem Kapitel wird ein RLDS-Schätzer entwickelt, der die Verständlichkeit der prozessierten Sprachsignale erhöhen soll. Die Verbesserung der Sprachsignalverständlichkeit kann etwa den Nutzern von Hörgeräten zugute kommen [LFJA09]. Wie in Abschnitt 4.1 beschrieben, benötigt man dafür eine robuste und zugleich dynamische RLDS-Schätzung, die als Baustein 1 in einem System zur Sprachsignalentstörung wie in Abb. 2.2 eingesetzt werden kann. Laut den Untersuchungen in [KLHL09] verbessert eine Signalverarbeitung mit binären Filterfunktionen die Verständlichkeit der prozessierten Sprachsignale. Darauf beruht der Gedanke, eine dynamische RLDS-Schätzung zu entwickeln, die über das *a posteriori* SNR zum schnelleren Schalten zwischen dem minimalen und dem maximalen Wert die Filterfunktion beiträgt. Da von den RLDS-Schätzern, die in Abschnitt 3.1 vorgestellt wurden, laut Tab. 3.1 nur das *Minimum Statistics* Verfahren aus [Mar01] auf die Technik der ausgangsseitigen Glättung verzichtet, ist es prädestiniert, den gestellten Anforderungen zu genügen. Darauf deuten auch die subjektiven und objektiven Tests in [Mar01] hin, wo neben verbesserter Signalqualität auch höhere Verständlichkeit der prozessierten Signale notiert wird.

Einen wesentlichen Beitrag zu den hervorragenden Eigenschaften des MS-Verfahrens leistet die recheneffiziente Minimumsuche über das geglättete Spektrogramm, das mit Hilfe einer optimierten rekursiven Glättung erster Ordnung realisiert wird. Ein zeitvarianter Glättungsparameter wird dabei im MMSE-Sinne optimal gewählt und über eine Steuerungsfunktion berechnet, die in den Experimenten mit Sprachsignalen allerdings einige Defizite aufweist [Mar01]. Und obwohl diese über eine heuristische Anpassung zum Teil aufgefangen werden, leidet die Steuerungsfunktion immer noch an mangelnder Verfolgungsfähigkeit in den Zeit-Frequenz-Punkten mit einem niedrigen lokalen SNR, wie in weiteren Untersuchungen gezeigt wird. Diese Beobachtungen dienen als Motivation für die Suche nach einer alternativen Steuerungsfunktion, die in diesem Kapitel vorgestellt wird.

Dafür wird zunächst in Abschnitt 5.1 die Herleitung der herkömmlichen Steuerungsfunktion des MS-Verfahrens aus [Mar01] rekapituliert. Im Abschnitt 5.2 wird basierend auf einer alternativen statistischen Modellierung zunächst eine Bayessche Glättung eingeführt, die zu einem im Bayesschen Sinne optimalen Glättungsparameter führt. Anschließend wird eine alternative Steuerungsfunktion vorgestellt, welche die konventionelle Steuerungsfunktion des MS-Verfahrens ersetzen kann. Im Abschnitt 5.3 wird die alternative Steuerungsfunktion in den experimentellen Untersuchungen mit den Sprachsignalen optimiert, die von weißem Rauschen gestört werden. Ergebnisse der umfangreichen Untersuchungen mit Sprachsignalen verschiedener Datenbanken, die mit unterschiedlichen Rauschtypen gestört werden, werden in Abschnitt 5.4 beschrieben. Man beachte, dass das in diesem Kapitel zu entwickelnde Verfahren zur RLDS-Schätzung zum ersten Mal in [CHU15] vorgestellt wurde.

5.1. Optimale Glättung des *Minimum Statistics* Verfahrens

Basierend auf der Additivität (2.3) und der statistischen Modellierung (2.4) und (2.5) mit den Definitionen (2.6) und (2.7) unterliegt die momentane spektrale Leistung $|Y(k,\ell)|^2$ des gestörten Sprachsignals einer Exponentialverteilung $p_{|Y(k,\ell)|^2}(y) = \text{Exp}(y; \lambda_Y(k,\ell))$ mit der Verteilungsdichtefunktion

$$\text{Exp}(y; \lambda_Y(k,\ell)) = \frac{1}{\lambda_Y(k,\ell)} \cdot \exp\left(-\frac{y}{\lambda_Y(k,\ell)}\right) \cdot U(y), \tag{5.1}$$

wobei $\lambda_Y(k,\ell) = \lambda_S(k,\ell) + \lambda_D(k,\ell)$ und $U(y)$ jeweils das RLDS eines gestörten Sprachsignals und die Einheitssprungfunktion (engl. *unit step function*) sind [Mar01]. Eine kausale RLDS-Schätzung aus den vergangenen Werten von $|Y(k,\ell)|^2$ ist besonders herausfordernd in den Zeit-Frequenz-Punkten mit Sprachsignalpräsenz und in Gegenwart einer nichtstationären Störung. Auf der anderen Seite ist sie relativ einfach zu realisieren in den Zeit-Frequenz-Punkten mit Sprachsignalabwesenheit $\lambda_Y(k,\ell) = \lambda_D(k,\ell)$ und in Gegenwart einer Störung, die in einer gewissen Zeitspanne (z. B. von L STFT-Rahmen) als ein stationärer Zufallsprozess $\lambda_D(k,\ell) = \lambda_D(k)$ modelliert werden kann. In diesem Fall könnte die RLDS-Schätzung über einen biasfreien ML-Schätzer $\hat{\lambda}_D^{\text{ML}}(k) = (1/L) \cdot \sum_{\ell=1}^{L} |Y(k,\ell)|^2$ berechnet werden, der in einer nichtrekursiven arithmetischen Mittelung resultiert.

Obwohl bei einer ML-Schätzung die zu schätzende Größe als ein unbekannter Parameter betrachtet wird, ist der ML-Schätzer $\hat{\lambda}_D^{\text{ML}}(k)$ selbst eine Zufallsvariable, deren Verteilungsdichtefunktion $p_{\hat{\lambda}_D^{\text{ML}}(k,\ell)}(\lambda) = \chi^2(\lambda;\, \nu, \lambda_D)$ laut [Mar94,ML01] asymptotisch einer skalierten Chi-Quadrat (χ^2) Verteilung unterliegt:

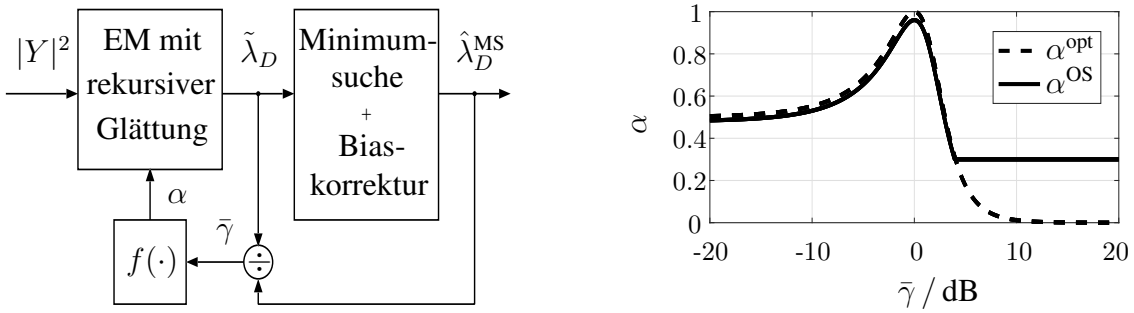
$$\chi^2(x;\, \nu, \lambda_D) = \left(\frac{\nu}{2\lambda_D}\right)^{\nu/2} \cdot \frac{x^{\nu/2-1}}{\Gamma(\nu/2)} \cdot \exp\left(-\frac{\nu}{2\lambda_D} \cdot x\right) \cdot U(x), \tag{5.2}$$

wobei $\nu \in \mathbb{R}_{>0}$ und $\Gamma(\cdot)$ jeweils ein Freiheitsgrad und die Gamma-Funktion sind. Die wahre RLDS λ_D fließt in die χ^2-Verteilung als ein weiterer Parameter ein und bestimmt somit auch die Schätzfehlervarianz $\text{var}(\hat{\lambda}_D^{\text{ML}}(k)) = 2\lambda_D^2/\nu$. Würden $Y(k,\ell)$ als unkorrelierte Zufallsvariablen betrachtet, wäre der Freiheitsgrad $\nu = L$ für $k \in \{0,\, k_{\text{Nyq}}\}$ und $\nu = 2L$ sonst für $k \in \{1, \ldots, k_{\text{Nyq}}\text{-}1\}$ zu wählen sein. Durch die Signalverarbeitung im STFT-Bereich mit den überlappenden Rahmen korrelieren allerdings die benachbarten Zufallsvariablen $Y(k,\ell)$ mit positiven Kovarianzen, sodass die Varianz des ML-Schätzers wächst und ein angepasster Freiheitsgrad $\nu_{\text{adj}} = \nu/a$ verwendet werden muss, der mit Hilfe eines Korrekturfaktors $a > 1$ berechnet werden kann [Mar94].

In einer Realisierung wird die nicht rekursive arithmetische Mittelung durch eine rekursive Filterung erster Ordnung ersetzt, die in einem geglätteten Spektrogramm resultiert

$$\tilde{\lambda}_D(k,\ell) = \alpha_{\text{MS}}(k,\ell) \cdot \tilde{\lambda}_D(k,\ell-1) + (1 - \alpha_{\text{MS}}(k,\ell)) \cdot |Y(k,\ell)|^2, \tag{5.3}$$

wobei $\alpha_{\text{MS}}(k,\ell)$ ein zeitvarianter Glättungsfaktor des MS-Verfahrens ist. Diese Art der Glättung begegnet uns in vielen RLDS-Schätzern. Während die Verfahren aus [Mar01, CB02, Coh03, KPC09] von einem zeitvarianten Glättungsparameter Gebrauch machen, verwenden RLDS-Schätzer aus [Hir93, Mar94, Dob95, Yu09] einen zeitinvarianten (konstanten) Glättungsparameter $\alpha(k,\ell) = \alpha$. Ungeachtet einer anderen Berechnungsweise wird $\tilde{\lambda}_D(k,\ell)$



(a) Steuerungsfunktion im MS-Verfahren

(b) Steuerungsfunktion $\alpha = f(\bar{\gamma})$

Abbildung 5.1.: Eingangsseitige rekursive Glättung in *Minimum Statistics* Verfahren mit der konventionellen Steuerungsfunktionen $\alpha = f(\bar{\gamma})$.

ähnlich wie $\hat{\lambda}_D^{\text{ML}}(k)$ approximativ mit einer χ^2-Verteilung modelliert mit einem Freiheitsgrad ν, der von α abhängig ist [Mar94].

Das geglättete Spektrogramm $\tilde{\lambda}_D(k,\ell)$ dient im MS-Verfahren als Grundlage für die darauf folgende Minimumsuche. Das gefundene Minimum wird dann einer Biaskorrektur unterzogen, die im RLDS-Schätzwert des MS-Verfahrens $\hat{\lambda}_D^{\text{MS}}(k,\ell)$ resultiert. Das Zusammenspiel verschiedener Komponenten des MS-Schätzers ist in Abb. 5.1 (a) dargestellt. Hier ist auch eine Steuerungsfunktion $f(\cdot)$ zu sehen, mit deren Hilfe der Glättungsparameter $\alpha_{\text{MS}}(k,\ell)$ berechnet wird. Vor der Verwendung in (5.3) wird dieser allerdings einer sogenannten Fehlerüberwachung (engl. *error monitoring*, EM) unterzogen, welche in [Mar01] mit Hilfe der Gleichungen (9)-(11) realisiert und hier nicht näher erläutert wird.

Bei der Herleitung der Steuerungsfunktion wird in [Mar01] zunächst eine optimale Steuerungsfunktion $f_{\text{opt}}(\cdot)$ unter der Annahme der Sprachsignalabwesenheit $Y(k,\ell) = D(k,\ell)$ hergeleitet, indem folgende MSE-Kostenfunktion minimiert wird

$$\alpha^{\text{opt}}(k,\ell) = \underset{\alpha}{\text{argmin}}\ \mathbb{E}\left[\left(\tilde{\lambda}_D(k,\ell) - \lambda_D(k,\ell)\right)^2 \middle|\ \tilde{\lambda}_D(k,\ell-1)\right]. \tag{5.4}$$

Die Minimierung von (5.4) mit der Rekursion (5.3) führt zur optimalen Steuerungsfunktion:

$$\alpha^{\text{opt}}(k,\ell) = \frac{1}{1 + (\bar{\gamma}(k,\ell-1) - 1)^2}, \tag{5.5}$$

wobei $\bar{\gamma}(k,\ell-1) = \tilde{\lambda}_D(k,\ell-1)/\lambda_D(k,\ell)$ eine geglättete *a posteriori* SNR ist, bei dem in einer Realisierung $\lambda_D(k,\ell)$ durch den letzten MS-Schätzwert $\hat{\lambda}_D^{\text{MS}}(k,\ell-1)$ ersetzt wird:

$$\bar{\gamma}(k,\ell-1) = \frac{\tilde{\lambda}_D(k,\ell-1)}{\hat{\lambda}_D^{\text{MS}}(k,\ell-1)}. \tag{5.6}$$

Man beachte, dass die Glättung (5.3) bei gestörten Sprachsignalen dem Verlauf von $|Y(k,\ell)|^2$ auch in der Sprachsignalpräsenz schnell folgen soll.

Jedoch in einer Realisierung führt die optimale Steuerungsfunktion $\alpha^{\text{opt}} = f_{\text{opt}}(\bar{\gamma})$, die in Abb. 5.1 (b) über $\bar{\gamma}_{\text{dB}} = 10 \cdot \log_{10} \bar{\gamma}$ dargestellt ist, zur folgenden Defiziten im geglätteten Spektrogramm $\tilde{\lambda}_D(k,\ell)$:

- **Totaler Stillstand**: Wie in [Mar01] erwähnt, ist ein totaler Stillstand (engl. *deadlock*) von $\tilde{\lambda}_D(k,\ell)$ und infolgedessen auch von $\hat{\lambda}_D^{\text{MS}}(k,\ell)$ für $\bar{\gamma} \approx 1$ möglich. Denn für ein relativ breites Intervall von $\bar{\gamma}$ um den Wert $\bar{\gamma} = 1$ liefert die optimale Steuerungsfunktion einen verhältnismäßig großen Glättungsparameter $\alpha^{\text{opt}} \approx 1$. In diesem Fall kann das geglättete Spektrogramm $\tilde{\lambda}_D(k,\ell)$ sich laut (5.3) kaum verändern und ist somit gefährdet, in einem konstanten Schätzwert stecken zu bleiben.

- **Hohe Schätzfehlervarianz**: Zum anderen erzeugen die sehr kleinen Werte des Glättungsparameters $\alpha^{\text{opt}} \to 0$ für $\bar{\gamma} \gg 1$ eine hohe Schätzfehlervarianz in $\hat{\lambda}_D^{\text{MS}}(k,\ell)$, wie in [Mar01] berichtet wird. Dieses Defizit ist allerdings durch Verwendung einer geglätteten Referenz für die wahre RLDS $\lambda_D(k,\ell)$ wie in (2.48) begründet, welche die kleinen Werte von α^{opt} härter bestraft. An sich sind die kleinen Werte von α^{opt} in der Sprachsignalpräsenz, die für $\bar{\gamma} \gg 1$ gegeben ist, von Vorteil, denn so kann das Einsetzen eines Sprachsignals (engl. *speech onsets*) und sein Aufhören von $\tilde{\lambda}_D(k,\ell)$ besser verfolgt werden.

- **Mangelnde Verfolgungsfähigkeit**: Ein Defizit der optimalen Steuerungsfunktion, das in [Mar01] allerdings unbeachtet bleibt, ist eine mangelnde Verfolgungsfähigkeit (engl. *tracking ability*) von $\tilde{\lambda}_D(k,\ell)$ für $\bar{\gamma} < 1$, wo der Glättungsparameter auf den theoretischen minimalen Wert $\lim_{\gamma\to 0} \alpha^{\text{opt}} = 0.5$ begrenzt ist. Dieser resultiert daraus, dass das MSE-Maß aus (5.4) die Unterschätzung weniger bestraft als die Überschätzung. Somit wird die Verfolgungsfähigkeit der rekursiven Glättung (5.3) in den Zeitspannen eingeschränkt, wo bei schneller Abnahme von $|Y(k,\ell)|^2$ das geglättete Spektrogramm $\tilde{\lambda}_D(k,\ell)$ unter das Niveau von $\hat{\lambda}_D^{\text{MS}}(k,\ell)$ absinkt. Dies kann die Minimumsuche des MS-Verfahrens beeinträchtigen, die auf die Glättung (5.3) folgt. Durch die Untergrenze von α^{opt} bekommt die Verfolgungsgeschwindigkeit v_P der rekursiven Glättung eine Obergrenze, dessen Wert laut (A.19) von der Abtastrate F_s und vom STFT-Rahmenvorschub R abhängt.

Um die ersten beiden Defizite der optimalen Steuerungsfunktion $f_{\text{opt}}(\bar{\gamma})$ zu mildern, wird sie in [Mar01] zur *optimally smoothed* (OS) Steuerungsfunktion $f_{\text{OS}}(\bar{\gamma})$ heuristisch angepasst, die im MS-Verfahren tatsächlich zum Einsatz kommt. Um den totalen Stillstand abzuschwächen, wird zum einen ein maximaler Glättungsparameter $\alpha_{\text{max}} = 0.96$ eingeführt, der bei den in [Mar01] gewählten Parametern $F_s = 8\,\text{kHz}$ und $R = 2^7$ dafür sorgt, dass die Verfolgungsfähigkeit der rekursiven Glättung oberhalb von etwa $v_P = 11.1\,\text{dB/s}$ bleibt. Zum anderen wird eine Untergrenze $\alpha_{\text{min}} = 0.3$ deklariert, die bei $\bar{\gamma} \approx 4\,\text{dB}$ eingreift und eine hohe Schätzfehlervarianz verhindert. α_{min} in [Mar01] sorgt für eine Obergrenze der Verfolgungsfähigkeit der rekursiven Glättung, die bei etwa $v_P = 327\,\text{dB/s}$ liegt. Somit wird der Glättungsparameter α^{OS} des MS-Verfahrens folgendermaßen berechnet

$$\alpha^{\text{OS}}(k,\ell) = \max\left(\alpha_{\text{max}} \cdot \alpha^{\text{opt}}(k,\ell),\ \alpha_{\text{min}}\right). \tag{5.7}$$

In Abb. 5.1 (b) ist neben der optimalen Steuerungsfunktion α^{opt} aus (5.5) als Funktion von $\bar{\gamma}$ auch α^{OS} aus (5.7) dargestellt.

Außerdem wird in [Mar01] vorgeschlagen, die Untergrenze α_{min} in Abhängigkeit vom globalen eingangsseitigen SNR_{IN} aus (2.40) zu berechnen:

$$\alpha_{\text{min}}(\text{SNR}_{\text{IN}}) = \min\left(0.3\,,\ \text{SNR}_{\text{IN}}^{-\frac{R}{0.064\cdot F_s}}\right), \tag{5.8}$$

$\bar{\gamma}$	Eigenschaften des MS-Schätzers	$\alpha^{\text{opt}}(\bar{\gamma})$	$\alpha^{\text{OS}}(\bar{\gamma})$
< 1	Mangelnde Verfolgungsfähigkeit	☹	☹
≈ 1	Totaler Stillstand	☹	☺
$\gg 1$	Hohe Schätzfehlervarianz	☹	☺
Zusätzliche Parameter		☺	☹

Tabelle 5.1.: Eigenschaften des MS-Verfahrens bei Verwendung von Steuerungsfunktionen $\alpha^{\text{opt}}(\bar{\gamma})$ aus (5.5) und $\alpha^{\text{OS}}(\bar{\gamma})$ aus (5.7) mit $\alpha_{\text{max}} = 0.96$ und $\alpha_{\text{min}} = 0.3$.

wobei SNR_{IN} ein absoluter Wert des eingangsseitigen SNR (also nicht in dB gemessen) ist [Mar01], vergleiche mit (A.20) in Anhang A. Die Gleichung (5.8) ermöglicht die Werte $\alpha_{\text{min}} < 0.3$ für die Werte von $\text{SNR}_{\text{IN}} > 21\,\text{dB}$. Allerdings muss dafür in einer Realisierung SNR_{IN} zusätzlich geschätzt werden.

Die Eigenschaften des MS-Verfahrens bei Verwendung von Steuerungsfunktionen $\alpha^{\text{opt}}(\bar{\gamma})$ aus (5.5) und $\alpha^{\text{OS}}(\bar{\gamma})$ aus (5.7) mit $\alpha_{\text{max}} = 0.96$ und $\alpha_{\text{min}} = 0.3$ sind in Tab. 5.1 aufgelistet. Hier wird ersichtlich, dass die Milderung von Defiziten der optimalen Steuerungsfunktion auf Kosten von zwei zusätzlichen heuristisch eingeführten Parametern stattfindet, die nicht für alle Szenarien optimal sein können. Außerdem führt die OS-Steuerungsfunktion α^{OS} immer noch zur mangelnden Verfolgungsfähigkeit der rekursiven Glättung. Diese Beobachtungen dienen als Motivation für die Suche nach einer alternativen Steuerungsfunktion.

5.2. Glättung mit einer alternativen Steuerungsfunktion

Im Unterschied zum Ansatz in [Mar01], in dem das wahre RLDS $\lambda_D(k,\ell)$ als ein unbekannter Parameter behandelt wird, wird es hier als eine Zufallsvariable modelliert. Darauf basierend können Bayessche Schätzer von $\lambda_D(k,\ell)$ für den Fall einer stationären Störung hergeleitet werden, die interessanterweise in einer rekursiven Glättung münden. Während in [Mar01] die rekursive Glättung heuristisch angenommen wird, ergibt sie sich mit dem neuen Ansatz allein aus der stochastischen Inferenz. Dabei bekommt der Glättungsparameter eine im Bayesschen Sinne optimale Form, die für die Anwendung bei einer nichtstationären Störung erweitert werden kann. Dadurch entsteht eine alternative Steuerungsfunktion, die im MS-Schätzer eingesetzt wird und in den experimentellen Untersuchungen zu höheren STOI-Werten führt.

Da in einer Realisierung weder das wahre RLDS $\lambda_D(k,\ell)$ noch seine Verteilungsdichte $p_{\Lambda_D}(\lambda_D)$ bekannt sind, hat man bei der statistischen Modellierung der Zufallsvariable $\Lambda_D(k,\ell)$ gewisse Freiheiten. Eine Abhilfe bei der Suche nach einer geeigneten Verteilung schafft die Beobachtungsverteilung $p_{|Y(k,\ell)|^2}(y) = \text{Exp}(y \,|\, \lambda_D(k,\ell))$ aus (5.1) für die Annahme der Sprachsignalabwesenheit[1]. Um die bevorstehende Bayessche Schätzung recheneffizient gestalten zu können, wird das Konzept der konjugierten *a priori* Verteilung herangezogen und $p_{\Lambda_D(k,\ell)}(\lambda_D)$ zu einer inversen skalierten χ^2 (Inv-χ^2) Verteilung gewählt:

$$\text{Inv-}\chi^2(x;\,\nu,\tau^2) = \left(\frac{\nu\tau^2}{2}\right)^{\nu/2} \cdot \frac{x^{-\nu/2-1}}{\Gamma(\nu/2)} \cdot \exp\left(-\frac{\nu\tau^2}{2}\cdot\frac{1}{x}\right) \cdot U(x), \tag{5.9}$$

[1] Statt des Semikolons wie in (5.1) wird hier ein senkrechter Strich verwendet, um zu verdeutlichen, dass $\lambda_D(k,\ell)$ hier als eine Zufallsvariable modelliert wird und nicht mehr als ein Parameter.

wobei $\nu \in \mathbb{R}_{>0}$ und $\tau^2 \in \mathbb{R}_{>0}$ jeweils ein Freiheitsgrad und ein Skalierungsparameter sind[2].

Gegeben eine Beobachtung im aktuellen Zeit-Frequenz-Punkt $|Y(\ell)|^2 \triangleq |Y(k,\ell)|^2$ und die *a priori* Verteilungsdichtefunktion $p_{\Lambda_D(\ell-1)}(\lambda_D) = \text{Inv-}\chi^2(\lambda_D;\, \nu(\ell-1), \tau^2(\ell-1))$ des vorigen Rahmens, ist die *a posteriori* Verteilungsdichtefunktion des aktuellen Rahmens, berechnet für die Sprachsignalabwesenheit über die Bayessche Regel für bedingte Wahrscheinlichkeiten, auch eine skalierte inverse χ^2-Verteilung $p_{\Lambda_D(\ell)}(\lambda_D) = \text{Inv-}\chi^2(\lambda_D;\, \nu(\ell), \tau^2(\ell))$ mit den Parametern

$$\nu(\ell) = \nu(\ell-1) + 2 \tag{5.10}$$

$$\tau^2(\ell) = \frac{\nu(\ell-1)}{\nu(\ell)} \cdot \tau^2(\ell-1) + \frac{2}{\nu(\ell)} \cdot |Y(\ell)|^2. \tag{5.11}$$

Da man in einer Realisierung jedoch immer an einem bestimmten RLDS-Schätzwert interessiert ist, soll ein konkreter Punktschätzer, wie ein MMSE- oder ein MAP-Schätzer, gewählt werden, die bei einer inversen skalierten Chi-Quadrat-Verteilung wie folgt berechnet werden:

$$\tilde{\lambda}_D^{\text{MMSE}} = \frac{\nu}{\nu-2} \cdot \tau^2 \qquad (5.12) \qquad\qquad \tilde{\lambda}_D^{\text{MAP}}(\ell) = \frac{\nu}{\nu+2} \cdot \tau^2. \qquad (5.13)$$

Man beachte, dass der MMSE-Schätzer aus (5.12) nur für $\nu > 2$ gültig ist. Um sich für einen bestimmten Bayesschen Schätzer zu einem späteren Zeitpunkt entscheiden zu können, wird basierend auf (5.12) und (5.13) ein verallgemeinerter Punktschätzer definiert

$$\tilde{\lambda}_D = \frac{\nu}{\nu+\Delta\nu} \cdot \tau^2, \tag{5.14}$$

der für $\Delta\nu = -2$ zu einem MMSE-Schätzer und für $\Delta\nu = 2$ zu einem MAP-Schätzer wird. Die Gültigkeit des verallgemeinerten Schätzers für den Fall $\Delta\nu < 0$ kann dann durch die Erfüllung der Bedingung $\nu > |\Delta\nu|$ gewährleistet werden.

Das Einsetzen von (5.10) und (5.14) in (5.11) resultiert in folgender Rekursionsgleichung:

$$\tilde{\lambda}_D(\ell) = \frac{\nu(\ell-1)+\Delta\nu}{\nu(\ell-1)+\Delta\nu+2} \cdot \tilde{\lambda}_D(\ell-1) + \frac{2}{\nu(\ell-1)+\Delta\nu+2} \cdot |Y(\ell)|^2, \tag{5.15}$$

die im Weiteren als eine Bayes-motivierte Glättung (engl. *Bayesian smoothing*, BS) bezeichnet wird. Ein Vergleich von (5.15) mit (5.3) führt zu einem Glättungsparameter

$$\alpha^{\text{BS}}(\ell) = \frac{\nu(\ell-1)+\Delta\nu}{\nu(\ell-1)+\Delta\nu+2}, \tag{5.16}$$

der eine Funktion von $\nu(\ell-1)$ und $\Delta\nu$ ist. Die Bedingung der Gültigkeit des MMSE-Schätzers sorgt dafür, dass der mit (5.16) berechnete Glättungsparameter $\alpha(\ell)$ im gewünschten Wertebereich $(0;\, 1)$ liegt. Dieser Sachverhalt wird in Abb. 5.2 (a) verdeutlicht, wo der Bayes-motivierte Glättungsparameter α als Funktion des Freiheitsgrades ν für unterschiedliche Werte von $\Delta\nu$ dargestellt ist. Hier wird ersichtlich, dass der gültige Glättungsparameter aus (5.16) alternativ auch folgendermaßen berechnet werden kann

$$\alpha^{\text{BS}}(\ell) = \max\left(\frac{\nu(\ell-1)+\Delta\nu}{\nu(\ell-1)+\Delta\nu+2},\, 0\right), \tag{5.17}$$

[2]Da in jedem Frequenzband k sowohl die statistische Modellierung als auch die Signalverarbeitung ähnlich ist, wird aus Gründen der Übersichtlichkeit der Frequenzindex k im Weiteren manchmal ausgelassen.

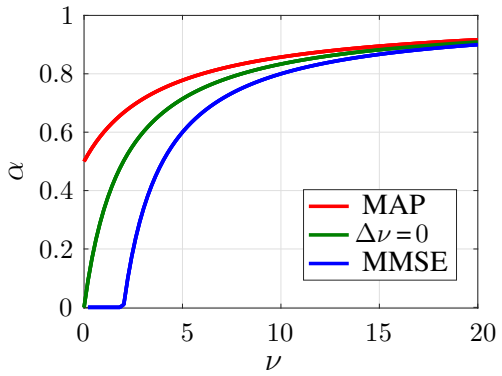
ohne dabei explizit zusätzliche Bedingungen für den Wertebereich von $\nu(\ell-1)$ angeben zu müssen. Für die Untergrenze des Glättungsparameters α^{BS} gilt dann

$$\alpha_{\min}^{\mathrm{BS}} = \lim_{\nu \to 0} \alpha^{\mathrm{BS}}(\nu) = \max\left(\frac{\Delta\nu}{\Delta\nu + 2},\, 0\right). \tag{5.18}$$

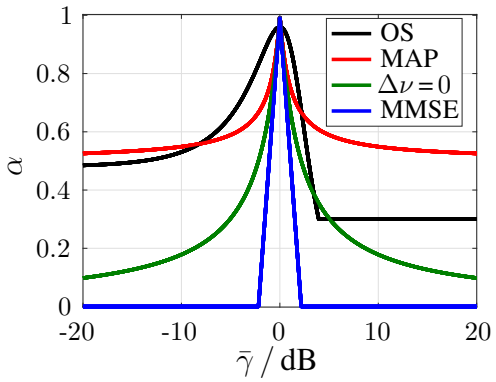
Daraus wird ersichtlich, dass neben den konventionellen MAP- und MMSE-Schätzwerten auch andere Punktschätzer durchaus betrachtet werden können. So führt die Verwendung von $\Delta\nu = 0$ zu einem Punktschätzer, der in dem Fall der inversen skalierten Chi-Quadrat-Verteilung genau in der Mitte zwischen dem MAP- und dem MMSE-Schätzwert liegen. Der Glättungsparameter $\alpha^{\mathrm{BS}}(\nu)$ ist für den Fall $\Delta\nu = 0$ in Abb. 5.2 (a) auch dargestellt.

Unabhängig vom gewählten $\Delta\nu$ zeigen die Kurven in Abb. 5.2 (a) allerdings, dass der Glättungsparameter α^{BS} mit steigendem Freiheitsgrad ν auch größer gewählt wird. Bei den großen Werten des Glättungsparameters $\alpha^{\mathrm{BS}} \to 1$ geht die aktuelle Beobachtung $|Y(\ell)|^2$ mit einem kleinen Gewicht in die Bayes-motivierten Glättung (5.15) ein, denn der Bayessche Schätzer verlässt sich zunehmend auf das *a priori* Wissen aus dem vorigen Rahmen. Dies würde im Fall eines stationären Rauschens sicherlich ein wünschenswertes Verhalten sein. Im Gegenzug gehen kleinere Werte von ν mit kleinen Werten des Glättungsparameters einher, der bei den Punktschätzern mit $\Delta\nu < 0$ auch zu Null werden kann. In diesem Fall ignoriert der Bayessche Schätzer das *a priori* Wissen ganz und verlässt sich einzig und allein auf die aktuelle Beobachtung. Die Eigenschaft des Bayesschen Schätzers aus (5.15), die Glättung bei Bedarf komplett auszuschalten, unterscheidet ihn von den konventionellen Steuerungsfunktionen des MS-Verfahrens. Somit ist die erstrebenswerte Eigenschaft einer erhöhten Dynamik des RLDS-Schätzers gegeben.

Da sich der Glättungsparameter α^{BS} über seine Abhängigkeit von ν steuern lässt, fehlt nur noch ein letzter Schritt bis zur gesuchten alternativen Steuerungsfunktion. In diesem soll eine Funktion $\nu = g(\bar{\gamma})$ festgelegt werden, mit deren Hilfe der Freiheitsgrad $\nu(\ell-1)$ der *a priori* Verteilung $p_{\Lambda_D(\ell-1)}(\lambda_D)$ aus dem geglätteten *a posteriori* SNR $\bar{\gamma}$ aus (5.6) berechnet wird. Dabei muss die gesuchte Funktion eine Polstelle bei $\bar{\gamma} = 1$ besitzen, damit es an dieser Stelle



(a) Bayes-motivierte Glättungsparameter



(b) Neue Steuerungsfunktionen $\alpha^{\mathrm{BS}}(\bar{\gamma})$

Abbildung 5.2.: Bayes-motivierte Glättungsparameter $\alpha^{\mathrm{BS}}(\nu)$ aus (5.17) und die dazu gehörigen alternative Steuerungsfunktionen $\alpha^{\mathrm{BS}}(\bar{\gamma})$ aus (5.21) für solche Punktschätzer wie MAP ($\Delta\nu = 2$), MMSE ($\Delta\nu = -2$) und ein verallgemeinerter Punktschätzer für $\Delta\nu = 0$.

ähnlich wie bei der konventionellen Steuerungsfunktionen des MS-Verfahrens $\alpha^{\text{BS}} = 1$ gilt. Eine weitere wünschenswerte Eigenschaft wäre $\lim_{\bar{\gamma}\to 0} g(\bar{\gamma}) = \lim_{\nu\to\infty} g(\bar{\gamma}) = 0$, sodass $\lim_{\bar{\gamma}\to 0} \alpha^{\text{BS}} = \lim_{\nu\to\infty} \alpha^{\text{BS}} = \alpha^{\text{BS}}_{\text{min}}$ aus (5.18) erreicht wird. Aus diesen Überlegungen heraus wird vorgeschlagen, den Freiheitsgrad $\nu(\ell-1)$ aus (5.17) als eine Kehrwertfunktion

$$\nu(\ell-1) = \frac{1}{d_{\text{NS}}(\ell-1)} \tag{5.19}$$

eines Instationaritätsfaktors $d_{\text{NS}} \in \mathbb{R}_{>0}$ (engl. *degree of nonstationarity*) zu berechnen, der aus $\bar{\gamma}$ mit Hilfe des Betrages eines natürlichen Logarithmus wie folgt berechnet wird

$$d_{\text{NS}}(\ell-1) = |\ln \bar{\gamma}(\ell-1)|. \tag{5.20}$$

Die Inspiration für die Einführung des Instationaritätsfaktors stammt aus dem Beitrag [CK11], in dem mit dessen Hilfe ein Grad der Instationarität eines Sprachsignals gemessen wird. Und obwohl der Instationaritätsfaktor d_{NS} aus (5.20) anders als in [CK11] definiert wird, hat er dieselbe Aussagekraft. Wird $\bar{\gamma} \approx 1$ beobachtet, ist der Instationaritätsfaktor nahe Null. Entfernt sich $\bar{\gamma}$ von einer Eins, wächst der Instationaritätsfaktor und bleibt dabei dank der Betragsfunktion stets positiv. Erwähnenswert ist dabei noch, dass die logarithmischen Abstände vom Wert $\bar{\gamma} = 1$ gleich stark bestraft werden.

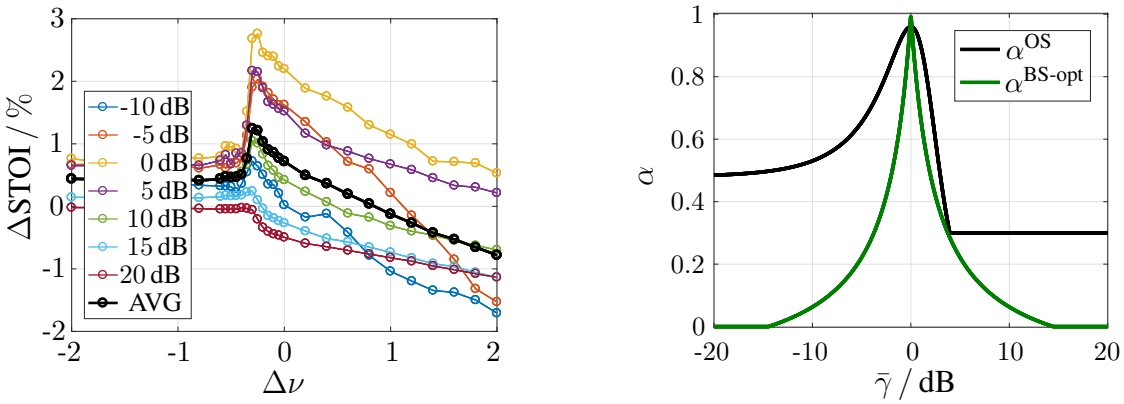
Einsetzen von (5.19) und (5.20) in (5.17) führt zur alternativen Steuerungsfunktion

$$\alpha^{\text{BS}}(\ell) = \max\left(\frac{1 + \Delta\nu \cdot |\ln \bar{\gamma}(\ell)|}{1 + (\Delta\nu + 2) \cdot |\ln \bar{\gamma}(\ell)|},\, 0\right) \tag{5.21}$$

die $\Delta\nu$ als einen Parameter hat. In Abb. 5.2 (b) wird die alternative Steuerungsfunktion für die Parameter $\Delta\nu = \{-2,\, 0,\, 2\}$ neben der konventionellen Steuerungsfunktion des MS-Verfahrens aus (5.5) und (5.7) über $\bar{\gamma}$ gemessen in Dezibel dargestellt. Das Problem des totalen Stillstandes wird dabei durch den Verlauf von $\alpha^{\text{BS}}(\bar{\gamma})$ um den Punkt $\bar{\gamma} = 1$ stark gemildert. Außerdem ist $\alpha^{\text{BS}}(\bar{\gamma})$ auf der logarithmischen Achse symmetrisch um den Wert $\bar{\gamma} = 1$, sodass die Bayes-motivierte Glättung (5.15) dem Spektrogramm des gestörten Sprachsignals nicht nur für $\bar{\gamma} > 1$ sondern auch für $\bar{\gamma} < 1$ bei Bedarf gleich schnell folgen kann. Eine entscheidende Rolle spielt dabei die Wahl des Parameters $\Delta\nu$, von dem ja nicht nur der minimale Wert des Glättungsparameters $\alpha^{\text{BS}}_{\text{min}}$ aus (5.18) abhängt, sondern auch der Verlauf der alternativen Steuerungsfunktion. Für negative Werte von $\Delta\nu$ bleibt α^{BS} positiv nur im Bereich $-\bar{\gamma}_0 < \bar{\gamma}_{\text{dB}} < \bar{\gamma}_0$ mit $\bar{\gamma}_0 = \frac{10}{|\Delta\nu|}\log_{10} e$ und nimmt sonst den Wert $\alpha^{\text{BS}} = 0$ an.

5.3. Datengetriebene Optimierung der alternativen Steuerungsfunktion

Da die alternative Steuerungsfunktion die objektive Sprachsignalverständlichkeit gemessen mit Hilfe des STOI-Maßes verbessern soll, wird der Parameter $\Delta\nu$ in den Experimenten mit Sprachsignalen optimiert, die mit einem additiven weißen Rauschen gestört sind. Dieses Simulationsszenario ist in der Sprachsignalverarbeitung typisch, wenn es um eine Analyse von Verfahren oder eine Parameteroptimierung geht [GBM08]. Dabei soll der optimale Wert von $\Delta\nu$ so gewählt werden, dass das STOI-Maß gemittelt über die typischen Werte des eingangsseitigen globalen SNR wie $\text{SNR}_{\text{IN}} \in \{-10, -5, 0, 5, 10, 15, 20\}$dB maximiert wird. In



(a) Optimum bei $\Delta\nu_{\text{opt}} = -0.3$

(b) $\alpha^{\text{OS}}(\bar{\gamma})$ und optimale $\alpha^{\text{BS-opt}}(\bar{\gamma})$

Abbildung 5.3.: Optimierung der alternativen Steuerungsfunktion in den Experimenten mit weißem Rauschen.

den Experimenten werden dabei die ungestörten Sprachsignale der TIMIT-Datenbank verwendet, die eine Abtastfrequenz von 16 kHz aufweisen. Für die Signalverarbeitung werden die relativ kurzen Sprachsignale der TIMIT-Datenbank zu längeren Sprachsignalen der Dauer 1 min für weibliche und männliche Sprecher getrennt zusammengesetzt. Somit wird die Notwendigkeit einer Anlaufphase vermieden, die sonst ein RLDS-Schätzer am Anfang einer jeden kurzen Äußerung bräuchte. Die ungestörten Sprachsignale werden anschließend mit dem weißen Rauschen der SPIB-Datenbank additiv überlagert, bevor sie einem System zur Sprachsignalentstörung wie in Abb. 2.2 dargeboten werden. Die Audiodaten der TIMIT- und SPIB-Datenbanken sind in Abschnitt 2.4 beschrieben.

Für die STFT wird die DFT-Länge $K = 2^9$, der Rahmenvorschub $R = 2^8$ und ein Hann-Fenster gewählt. Der zeitlicher Abstand zwischen den aufeinander folgenden STFT-Rahmen $\frac{R}{F_s} = 16\,\text{ms}$ entspricht somit dem aus [Mar01], sodass die verwendeten Glättungsparameter für die gleichen Zeitkonstanten und Verfolgungsgeschwindigkeiten der rekursiven Glättung (5.3) sorgen. Wie in [Mar01] wird die Minimumsuche des MS-Verfahrens über ein kausales Fenster der gesamten Länge $D_{\text{MS}} = U_{\text{MS}} \cdot V_{\text{MS}} = 96$ Rahmen, aufgeteilt in $U_{\text{MS}} = 8$ Unterfenstern der Länge $V_{\text{MS}} = 12$ Rahmen, durchgeführt[3]. In der alternativen Steuerungsfunktion (5.21) wird der Parameter $\Delta\nu$ im Bereich $[-2; 2]$ variiert. Als *a priori* SNR-Schätzer $\hat{\xi}(k, \ell)$ wird das *Decision-Directed* Verfahren aus [EM84] mit der Glättungskonstante $\alpha_{\text{DD}} = 0.95$ und der Untergrenze $\xi_{\text{min}} = -18\,\text{dB}$ verwendet. Als spektrale Filterfunktion $G(k, \ell)$ wird der OMLSA-Schätzer mit $G_{\text{min}} = -18\,\text{dB}$ aus [Coh01] eingesetzt.

In Abb. 5.3 (a) ist die Verbesserung des STOI-Maßes $\Delta\text{STOI} = \text{STOI}_{\text{OUT}} - \text{STOI}_{\text{IN}}$ dargestellt, die sich als Differenz von STOI-Werten am Ausgang des Systems STOI_{OUT} und an seinem Eingang STOI_{IN} darstellt. Die ΔSTOI-Werte sind über die simulierten Werte des Parameters $\Delta\nu$ für verschiedene SNR_{IN}-Werte aufgetragen. Außerdem ist die mittlere Verbesserung von ΔSTOI dargestellt, die als AVG bezeichnet wird. Wie die Experimente zei-

[3] Bezüglich der verwendeten *MATLAB*-Realisierung des MS-Verfahrens ist anzumerken, dass sie freundlicherweise von der Arbeitsgruppe des Autors des Verfahrens zur Verfügung gestellt wurde. In den experimentellen Untersuchungen wird in dieser Realisierung nur die Steuerungsfunktion ausgetauscht.

gen, hängt die Leistungsfähigkeit der alternativen Steuerungsfunktion stark von Werten von $\Delta\nu$ ab. Während der MAP-motivierte Glättungsparameter für $\Delta\nu = 2$ die Verständlichkeit der gestörten Sprachsignale insgesamt verschlechtert, gelingt es der MMSE-basierten Steuerungsfunktion für $\Delta\nu = -2$ einen kleinen positiven Wert ΔSTOI zu erreichen. Gemittelt über alle SNR_{IN}-Werte wird die größte Verbesserung der Sprachsignalverständlichkeit von etwa $1.24\,\%$ bei einem Wert von $\Delta\nu_{\text{opt}} = -0.3$ erreicht. Betrachtet man die resultierenden ΔSTOI-Werte für verschiedene Werte von SNR_{IN} fällt auf, dass die Verbesserung der Verständlichkeit bei den stark gestörten Sprachsignalen (ein im Praxis relevanter Fall) besonders groß ist. So wird bei $\text{SNR}_{\text{IN}} = 0\,\text{dB}$ alleine durch die Sprachsignalentstörung mit $\Delta\nu_{\text{opt}} = -0.3$ der STOI-Wert von $69.9\,\%$ um $2.68\,\%$ auf $72.58\,\%$ erhöht. Somit erweist sich ein verallgemeinerter Punktschätzer aus (5.14) für $\Delta\nu_{\text{opt}} = -0.3$ als optimal, der etwas näher zum MMSE- als zum MAP-Schätzer liegt. Man beachte, dass die experimentelle Untersuchungen mit allen 15 Rauschtypen der SPIB-Datenbank zum selben Optimum führten.

Die optimale Bayes-motivierte Steuerungsfunktion $\alpha^{\text{BS-opt}}$ aus (5.21) mit $\Delta\nu_{\text{opt}} = -0.3$ ist in Abb. 5.3 (b) neben der konventionellen OS-Steuerungsfunktion α^{OS} über $\bar{\gamma}$ in Dezibel dargestellt. Da der optimale Parameter $\Delta\nu_{\text{opt}}$ negativ ist, nimmt der Glättungsparameter α^{BS} die von Null unterschiedlichen Werte nur im Bereich $\bar{\gamma}_{\text{dB}} \in (-\bar{\gamma}_0;\, \bar{\gamma}_0)$ an, wobei $\bar{\gamma}_0 \approx 14.48\,\text{dB}$ gilt. Außerhalb dieses Bereichs für $|\bar{\gamma}_{\text{dB}}| > \bar{\gamma}_0$, wenn also $\bar{\gamma}$ sich zu weit von einer Eins entfernt, wird die Glättung (5.3) aufgrund von $\alpha^{\text{BS-opt}} = 0$ komplett ausgeschaltet, sodass das geglättete Spektrogramm der momentanen spektralen Leistung gleich gesetzt wird $\tilde{\lambda}_D(k,\ell) = |Y(k,\ell)|^2$. Die Verwendung des ungeglätteten Spektrogramms in der Minimumsuche des MS-Verfahrens scheint hinsichtlich der Sprachsignalverständlichkeit gemessen mit STOI-Maß unproblematisch zu sein. Im Unterschied zur Steuerungsfunktion α^{OS} macht $\alpha^{\text{BS-opt}}$ es der rekursiven Glättung (5.3) möglich, den ganzen Wertebereich der Verfolgungsgeschwindigkeiten $v_P \in (0;\, \infty)$ auszunutzen.

Um das Verhalten der beiden Steuerungsfunktionen aus Abb. 5.3 (b) im Anwendungsfall zu untersuchen, wird ein Experiment mit einem Störsignal in Abwesenheit eines Sprachsignals durchgeführt, d. h. $Y(k,\ell) = D(k,\ell)$. Als Störsignal wird dabei das nichtstationäre Rauschen *factory1* der SPIB-Datenbank verwendet. Im Rahmen dieses Experimentes wird aus dem Spektrogramm $|Y(k,\ell)|^2$ mit Hilfe des konventionellen MS-Verfahrens ein RLDS-Schätzwert $\hat{\lambda}_D^{\text{MS}}(k,\ell)$ berechnet, wie in Abb. 5.1 dargestellt. Die zeitlichen Verläufe des geglätteten Spektrogramms $\tilde{\lambda}_D^{\text{OS}}$ und des Glättungsparameters α^{OS} (nach der Durchführung der Fehlerüberwachung), die das MS-Verfahren dabei verwendet hat, sind in Abb. 5.4 dargestellt und zwar für das Frequenzband um $781,25$ Hz. Aus dem Schätzwert $\hat{\lambda}_D^{\text{MS}}(k,\ell)$ wird anschließend das geglättete SNR $\bar{\gamma}(k,\ell)$ bestimmt, mit dessen Hilfe der Glättungsparameter $\alpha^{\text{BS-opt}}$ über die optimale Bayes-motivierte Steuerungsfunktion aus (5.21) mit $\Delta\nu = -0.3$ berechnet wird. Zum Schluss wird der Glättungsparameter $\alpha^{\text{BS-opt}}$ der in Abb. 5.1 (a) bereits erwähnten Fehlerüberwachung unterzogen, bevor er für die Glättung des Spektrogramms in (5.3) eingesetzt wird. Die resultierenden $\tilde{\lambda}_D^{\text{BS-opt}}$ und $\alpha^{\text{BS-opt}}$ (nach der Durchführung der Fehlerüberwachung) sind in Abb. 5.4 ebenfalls dargestellt. Man beachte, dass sowohl die geglätteten Spektrogramme $\tilde{\lambda}_D^{\text{OS}}$ und $\tilde{\lambda}_D^{\text{BS-opt}}$ als auch die resultierenden Glättungsparameter α^{OS} und $\alpha^{\text{BS-opt}}$ basierend auf demselben Schätzwert $\hat{\lambda}_D^{\text{MS}}(k,\ell)$ berechnet werden.

Die in Abb. 5.4 dargestellte Zeitspanne wird in drei Bereiche aufgeteilt $\bar{\gamma} < 1$, $\bar{\gamma} > 1$ und $\bar{\gamma} \approx 1$, die gekennzeichnet sind. Eine bemerkenswerte Fähigkeit der Bayes-motivierten Steuerungsfunktion wird im ersten Bereich $\bar{\gamma} < 1$ dadurch ersichtlich, dass $\tilde{\lambda}_D^{\text{BS-opt}}$ dem zeitlichen Verlauf von $|Y(k,\ell)|^2 = |D(k,\ell)|^2$ schneller als $\tilde{\lambda}_D^{\text{OS}}$ folgen kann. Dies resultiert

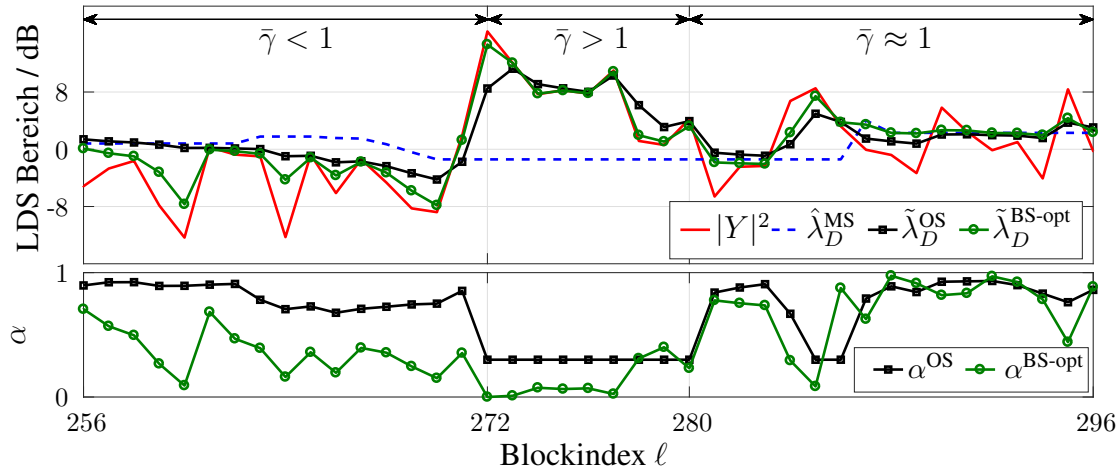


Abbildung 5.4.: Verläufe von den Größen aus Abb. 5.1 (a) im LDS-Bereich (oben) und die dazu gehörigen Glättungsparameter (unten) vorgefunden im Frequenzband um $781,25$ Hz.

aus den deutlich kleineren Werten von $\alpha^{\text{BS-opt}}$ hier im Vergleich zu den relativ großen Werten von α^{OS}, wie in der unteren Abbildung gut zu sehen ist. Die theoretische Untergrenze $\lim_{\bar{\gamma}\to 0} \alpha^{\text{OS}} = 0.48$, die in einer Obergrenze der Verfolgungsgeschwindigkeit von etwa $v_P \approx 199$ dB/s resultiert, ist der Grund hier für die mangelnde Fähigkeit des konventionellen MS-Verfahrens der momentanen spektralen Leistung schnell folgen zu können. Da momentane spektrale Leistung $|Y(k,\ell)|^2$ im Bereich $\bar{\gamma} < 1$ selbst bei Sprachsignalpräsenz vom Störsignal dominiert wird, soll die gute Verfolgungseigenschaft der Bayes-motivierten Steuerungsfunktion dem MS-Verfahren helfen, im Vergleich zur konventionellen Steuerungsfunktion ein besseres Minimum der spektralen Leistung des Störsignals zu finden.

Auch im Bereich $\bar{\gamma} > 1$ bleibt diese Überlegenheit erhalten, besonders wenn die momentane spektrale Signalleistung plötzlich steigt oder sinkt, was mit dem Verlauf der beiden Steuerungsfunktionen in Abb. 5.3 (b) einhergeht. Hier macht die heuristisch eingeführte Untergrenze $\alpha_{\text{min}} = 0.3$ dem MS-Verfahren das Leben schwer. In den Zeit-Frequenz-Punkten mit $\bar{\gamma} \approx 1$ verhalten sich die beiden Steuerungsfunktionen in etwa ähnlich, was im Fall einer stationären Störung auch erwünscht ist. Dabei benötigt die vorgeschlagene Steuerungsfunktion keinen zusätzlichen Parameter, denn das Problem des totalen Stillstandes, das bei α^{OS} durch die Einführung von $\alpha_{\text{max}} = 0.96$ vermieden wird, tritt hier nicht auf.

Da für eine gute Verständlichkeit der prozessierten Sprachsignale kleine Werte der Glättungsparameter erwünscht sind, wie die Optimierung der vorgeschlagenen Steuerungsfunktion zeigte, werden zusätzliche experimentelle Untersuchungen (wieder mit den Sprachsignalen gestört durch weißes Rauschen) mit dem Ziel durchgeführt, die konventionelle Steuerungsfunktion des MS-Verfahrens (5.7) hinsichtlich der Maximierung der STOI-Werte zu optimieren. Und tatsächlich, mit sinkenden Werten von α_{min} steigen die STOI-Werte der Ausgangssignale an, sodass $\alpha_{\text{min}} = 0$ sich als ein optimaler Parameter darstellt. Die Verwendung der im MMSE-Sinn optimalen Steuerungsfunktion (5.5), was der Wahl $\alpha_{\text{max}} = 1$ entsprechen würde, führte dagegen zur Verschlechterung der STOI-Werte. Aus diesen Gründen wird eine im STOI-Sinn optimale konventionelle Steuerungsfunktion für den Parametersatz $\alpha_{\text{max}} = 0.96$ und $\alpha_{\text{min}} = 0$ eingeführt, die im Weiteren als $\alpha^{\text{OS-opt}}$ bezeichnet wird. Wie man sich denken kann, führt die Verwendung von $\alpha^{\text{OS-opt}}$ zur Verbesserung der Verfolgungseigenschaft von α^{MS} nur im Bereich $\bar{\gamma} \gg 1$. Die Steuerungsfunktion $\alpha^{\text{OS-opt}}$ ermöglicht der

SNR, dB		-10	-5	0	5	10	15	20	AVG
STOI_{IN}		48.4	58.4	69.9	81.2	90.1	95.7	98.4	77.4
ΔSTOI	α^{OS}	-1.56	-3.67	-3.34	-2.21	-1.85	-1.64	-1.32	-2.23
	$\alpha^{\text{OS-opt}}$	-1.27	-1.94	-0.46	0.04	-0.28	-0.12	-0.15	-0.60
	$\alpha^{\text{BS-opt}}$	0.72	1.90	2.68	2.17	1.05	0.24	-0.06	1.24

Tabelle 5.2.: Verbesserung der Sprachsignalverständlichkeit gemessen in STOI für die Steuerungsfunktionen α^{OS}, $\alpha^{\text{OS-opt}}$ und $\alpha^{\text{BS-opt}}$ in den Experimenten mit weißem Rauschen.

rekursiven Glättung (5.3) bei den in [Mar01] verwendeten Parametern F_s und R die Verfolgungsgeschwindigkeiten im Bereich $v_P \in [11.1;\, \infty)$ dB/s.

Um die Leistungsfähigkeit der Steuerungsfunktionen zu vergleichen, werden in Tab. 5.2 neben den STOI_{IN}-Werten der gestörten Sprachsignale auch die Verbesserung der Sprachsignalverständlichkeit gemessen in $\Delta\text{STOI} = \text{STOI}_{\text{OUT}} - \text{STOI}_{\text{IN}}$ für drei Steuerungsfunktionen angegeben:

1. α^{OS}: konventionelle MS-Steuerungsfunktion aus (5.7) mit $\alpha_{\text{max}} = 0.96$ und $\alpha_{\text{min}} = 0.3$,
2. $\alpha^{\text{OS-opt}}$: optimale MS-Steuerungsfunktion aus (5.7) mit $\alpha_{\text{max}} = 0.96$ und $\alpha_{\text{min}} = 0$,
3. $\alpha^{\text{BS-opt}}$: optimale Bayes-motivierte Steuerungsfunktion aus (5.21) mit $\Delta\nu = -0.3$.

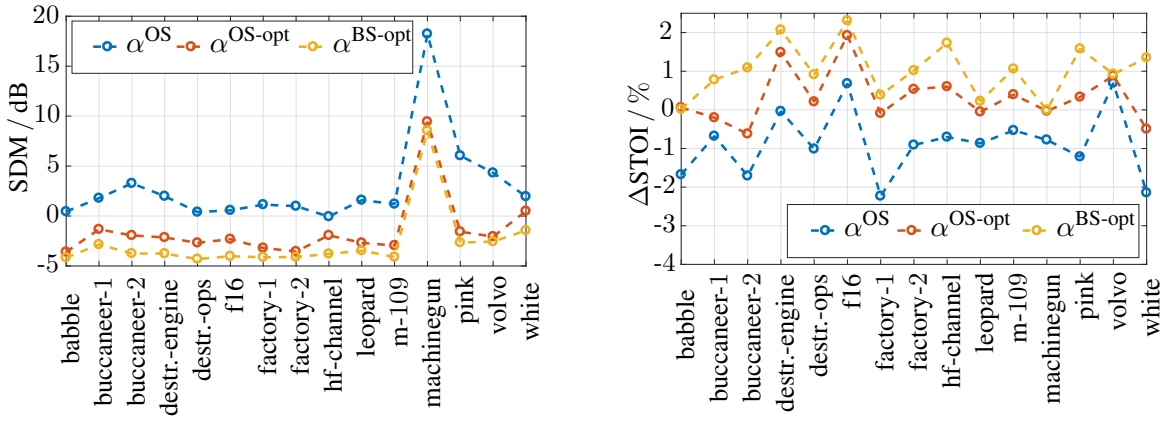
Wie man aus Untersuchungen wie in [HL07, LYZ+11] kennt, lässt sich die Verständlichkeit von gestörten Sprachsignalen durch die spektrale Entstörung mit konventionellen Algorithmen in der Regel nicht verbessern. Auch in unserem Fall bringt die spektrale Entstörung unter Verwendung der konventionellen Steuerungsfunktion α^{MS} durchgehend eine Verschlechterung der Sprachsignalverständlichkeit gemittelt über alle SNR_{IN}-Werte um insgesamt $-2.23\,\%$. Und obwohl die optimale MS-Steuerungsfunktion $\alpha^{\text{OS-opt}}$ im Vergleich zur konventionellen Steuerungsfunktion α^{OS} durchschnittlich einen um $1.63\,\%$ besseren STOI-Wert liefert, findet auch hier immer noch keine Verbesserung der Sprachsignalverständlichkeit statt. Erst beim Einsatz der optimalen Bayes-motivierten Steuerungsfunktion $\alpha^{\text{BS-opt}}$ werden positive ΔSTOI-Werte beobachtet, sodass durchschnittlich eine Verbesserung der Sprachsignalverständlichkeit von $77.4\,\%$ um $1.24\,\%$ auf $78.24\,\%$ festzustellen ist. Bemerkenswert ist, dass die ermittelte Verbesserung nur mit Hilfe der Optimierung eines einzigen Bausteins im System aus Abb. 2.2 erreicht wird.

5.4. Experimentelle Untersuchungen

Um die Leistungsfähigkeit verschiedener Steuerungsfunktionen in unterschiedlichen Simulationsszenarien zu untersuchen, werden zunächst umfangreiche Experimente mit den Störsignalen verschiedener Rauschtypen durchgeführt. Dafür werden die einminütigen ungestörten Sprachsignale der TIMIT-Datenbank aus Abschnitt 5.3 mit den Störsignalen aller 15 Rauschtypen der SPIB-Datenbank additiv überlagert und zwar bei den eingangsseitigen globalen SNR-Werten von $\text{SNR}_{\text{IN}} \in \{-10, -5, 0, 5, 10, 15, 20\}$dB. Anschließend werden die gestörten Sprachsignale durch die Verarbeitung im gleichen System entstört, das

in Abschnitt 5.3 bereits beschrieben wurde. Im MS-Verfahren werden dabei drei verschiedene Steuerungsfunktionen α^{OS}, $\alpha^{\text{OS-opt}}$ und $\alpha^{\text{BS-opt}}$ der Reihe nach eingesetzt. Bei diesen Experimenten wird nicht nur die Verständlichkeit der Sprachsignale in STOI sondern auch ihre Qualität in MOS-LQO$_{\text{WB}}$-Maß aus Abschnitt 2.3 gemessen. Außerdem wird die RLDS-Schätzung mit den LEM- und LEV-Maßen aus Abschnitt 2.3 bewertet, wobei das Spektrogramm des Störsignals $|D(k,\ell)|^2$ als RLDS-Referenz verwendet wird. Als weitere Maße für den mittleren RLDS-Schätzfehler werden die Itakura-Saito-Distanz (ISD) aus [GM76] und das SDM-Maß aus [IS06] berechnet. Im Unterschied zum LEM-Maß, das die Unterschätzung und die Überschätzung gleich bestraft, wird bei den ISD- und SDM-Maßen die Überschätzung stärker sanktioniert. Während das ISD-Maß alle Frequenzbänder gleichermaßen behandelt, ahndet das SDM-Maß die Überschätzung bei tiefen Frequenzen, wo die meiste Energie des Sprachsignals zu finden ist, stärker als die bei den hohen Frequenzen.

In Abb. 5.5 sind der mittlere SDM-Schätzfehler in Dezibel und die Verbesserung der Sprachsignalverständlichkeit ΔSTOI jeweils im Bild (a) und (b) für die Steuerungsfunktionen α^{OS}, $\alpha^{\text{OS-opt}}$ und $\alpha^{\text{BS-opt}}$ dargestellt, gemittelt über alle simulierten SNR$_{\text{IN}}$-Szenarien und über Signale von weiblichen und männlichen Sprechern. Wie in Abb. 5.5 (a) zu sehen ist, liefert die konventionelle Steuerungsfunktion α^{OS} den größten RLDS-Schätzfehler gemessen mit dem SDM-Maß. Die beiden anderen Steuerungsfunktionen $\alpha^{\text{OS-opt}}$ und $\alpha^{\text{BS-opt}}$ liefern einen deutlich kleineren Schätzfehler, wenn man die Ordinate in Dezibel in Betracht zieht. Von diesen schneidet allerdings $\alpha^{\text{BS-opt}}$ immer etwas besser ab als $\alpha^{\text{OS-opt}}$. Sonst fällt auf, dass der hoch nichtstationäre Rauschtyp *machinegun* dem MS-Schätzer große Schwierigkeiten bereitet und zwar unabhängig davon, welche Steuerungsfunktion verwendet wird. Für solche transienten Störungen können andere RLDS-Schätzer verwendet werden wie in [HDTC12, SMP13, TCG13], welche die besonderen Eigenschaften der transienten Störungen mit berücksichtigen. Betrachtet man die resultierenden ΔSTOI-Werte in Abb. 5.5 (b), stellt man fest, dass die konventionelle Steuerungsfunktion α^{OS} gemittelt über alle Rausch-



(a) Mittlerer Schätzfehler

(b) STOI-Verbesserung

Abbildung 5.5.: Mittlerer SDM-Schätzfehler und Verbesserung der Sprachsignalverständlichkeit gemessen in STOI (STOI$_{\text{IN}} = 77.25\,\%$ durchschnittlich) für drei Steuerungsfunktionen α^{OS}, $\alpha^{\text{OS-opt}}$ und $\alpha^{\text{BS-opt}}$ gemittelt über SNR$_{\text{IN}} \in \{-10, -5, 0, 5, 10, 15, 20\}$dB und über Signale von weiblichen und männlichen Sprechern.

-	LEM	ISD / dB	SDM	LEV	ΔMOS-LQO$_{\text{WB}}$	ΔSTOI / %
α^{OS}	**5.68**	41.6	1.65	**19.0**	**0.39**	-0.88
$\alpha^{\text{OS-opt}}$	8.65	24.9	0.61	40.3	0.18	0.36
$\alpha^{\text{BS-opt}}$	7.34	**21.2**	**0.45**	34.1	0.19	**1.10**

Tabelle 5.3.: Durchschnittliche Leistungsfähigkeit der Steuerungsfunktionen α^{OS}, $\alpha^{\text{OS-opt}}$ und $\alpha^{\text{BS-opt}}$ gemittelt über SNR$_{\text{IN}} \in \{-10, -5, 0, 5, 10, 15, 20\}$dB, über 14 Rauschtypen der SPIB-Datenbank (*machinegun* ausgenommen) und über Signale von weiblichen und männlichen Sprechern mit den durchschnittlichen MOS-LQO$_{\text{WB,IN}} = 1.40$ und STOI$_{\text{IN}} = 76.7\,\%$.

typen die Sprachsignalverständlichkeit um etwa $1\,\%$ des STOI-Maßes verschlechtert. Diese Lage verändert sich, wenn man die Steuerungsfunktion $\alpha^{\text{OS-opt}}$ verwendet und kleinere Werte des Glättungsparameters für $\bar{\gamma} \gg 1$ zulässt. Kommt allerdings die Steuerungsfunktion $\alpha^{\text{BS-opt}}$ zum Einsatz, ist eine weitere Verbesserung der Sprachsignalverständlichkeit zu notieren, die auf ein dynamisches Verhalten der geglätteten RLDS $\tilde{\lambda}_D(k, \ell)$ für $\bar{\gamma} \ll 1$ zurückzuführen ist. Bemerkenswert ist, dass die Steuerungsfunktion $\alpha^{\text{BS-opt}}$ für alle Rauschtypen (*babble* Rauschen ausgeschlossen[4]) durchgehend bessere STOI-Werte liefert.

Um den Einfluss verschiedener Steuerungsfunktionen auf die RLDS-Schätzung und die Eigenschaften der prozessierten Signale gemittelt über alle simulierten Szenarien zu untersuchen, werden die durchschnittlichen Werte aller berechneten Maße über die 14 Rauschtypen der SPIB-Datenbank in Tab. 5.3 angegeben. Dabei fließen die Maße, die sich für den Rauschtyp *machinegun* ergeben haben, in die Mittelung nicht ein, da dieser sich als eine transiente Störung erwies.

Beim Betrachten der resultierenden Maße fällt auf, dass die konventionelle Steuerungsfunktion α^{OS} den kleinsten LEM-Schätzfehler und die niedrigste Schätzfehlervarianz LEV liefert, die zur prozessierten Signalen mit bester Qualität gemessen in MOS-LQO$_{\text{WB}}$ führt. Die beiden anderen Steuerungsfunktionen verlieren zwar gegenüber der konventionellen Funktion in LEM-, LEV- und MOS-LQO$_{\text{WB}}$-Maßen deutlich, zeigen allerdings bessere ISD- und SDM-Werte, die sich letztendlich in der Verbesserung der Sprachsignalverständlichkeit widerspiegeln. Dabei deuten die kleinen ISD- und SDM-Werte bei den hohen LEM-Werten darauf hin, dass $\alpha^{\text{OS-opt}}$ und $\alpha^{\text{BS-opt}}$ die RLDS-Referenz seltener überschätzen, als dies bei α^{OS} der Fall ist. Hohe LEV-Werte reflektieren die Tatsache, dass die beiden optimierten Steuerungsfunktionen wie gewünscht zu dynamischen RLDS-Schätzwerten führen. Verglichen untereinander gewinnt bei den optimierten Steuerungsfunktionen die vorgeschlagene Funktion $\alpha^{\text{BS-opt}}$ gegenüber der Funktion $\alpha^{\text{OS-opt}}$ in allen berechneten Maßen. Insbesondere führt sie zu den Signalen mit deutlich besserer Sprachsignalverständlichkeit und übertrifft die konventionelle Steuerungsfunktion um gute $2\,\%$ des absoluten STOI-Maßes von durchschnittlich $75.8\,\%$ auf $77.8\,\%$. Sonst fällt auf, dass die Qualität und Verständlichkeit der prozessierten Sprachsignale wie auch in [HL07, LK11] nicht gleichzeitig verbessert werden können und eine gewisse Kompromissbereitschaft fordern.

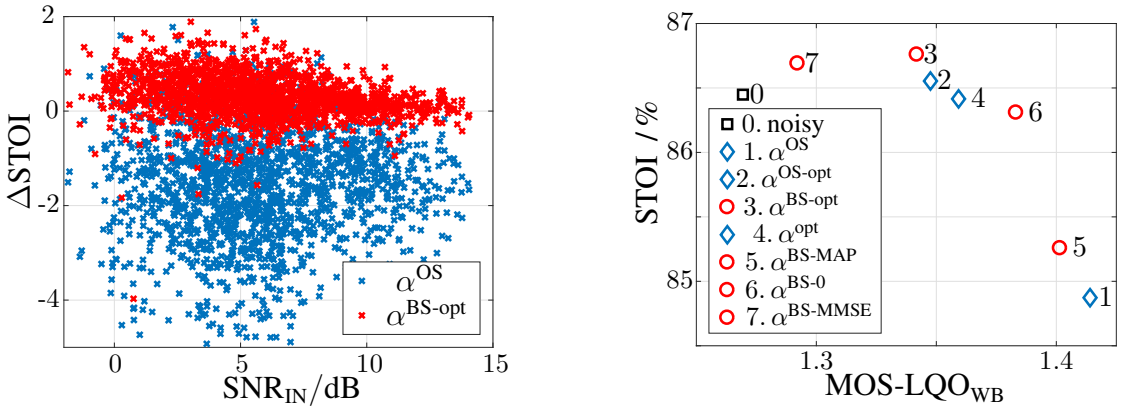
Um Leistungsfähigkeit der Steuerungsfunktionen auf einem ganz anderen Datensatz zu testen, werden experimentelle Untersuchungen auf den CHiME-3-Daten durchgeführt, die

[4] Die mittleren STOI-Werte der Steuerungsfunktion $\alpha^{\text{OS-opt}}$ sind beim *babble* Rauschen nur geringfügig besser als die der vorgeschlagenen Funktion $\alpha^{\text{BS-opt}}$.

in Abschnitt 2.4 beschrieben wurde. Dabei werden nur die Verständlichkeit und die Qualität der prozessierten Signale gemessen und anhand der STOI- und MOS-LQO$_{\text{WB}}$-Werte untersucht. In Abb. 5.6 (a) ist die Verbesserung der Verständlichkeit der prozessierten Signale ΔSTOI gegenüber den gestörten Eingangssignalen für alle verwendeten Audioaufnahmen der CHiME-3-Datenbank über das globale SNR der Eingangssignale SNR$_{\text{IN}}$ dargestellt, wenn die konventionelle Steuerungsfunktion α^{OS} und die vorgeschlagene Funktion $\alpha^{\text{BS-opt}}$ zum Einsatz kommen.

Wie man sieht, sind die CHiME-3-Daten stark verrauscht mit SNR$_{\text{IN}}$ im Bereich zwischen etwa -2 dB und 14 dB mit dem durchschnittlichen SNR$_{\text{IN}}$ von etwa 6 dB. Offenkundig liefert die vorgeschlagene Steuerungsfunktion $\alpha^{\text{BS-opt}}$ auch bei den CHiME-3-Daten bessere STOI-Werte als die konventionelle Funktion α^{OS} und verschiebt die meisten ΔSTOI-Werte vom negativen Wertebereich in den positiven. Während die konventionelle Steuerungsfunktion gemittelt über alle Audioaufnahmen den STOI-Wert von $84.9\,\%$ erreicht, erhöht die vorgeschlagene Steuerungsfunktion die Sprachsignalverständlichkeit auf den durchschnittlichen STOI-Wert von $86.8\,\%$. Ähnlich wie bei den Experimenten mit den Daten der TIMIT- und SPIB-Datenbanken wird auch bei den CHiME-3-Daten beim Verwenden von $\alpha^{\text{BS-opt}}$ statt α^{OS} die Verbesserung der Sprachsignalverständlichkeit von etwa $2\,\%$ des absoluten STOI-Maßes beobachtet. Außerdem bestätigt diese Untersuchung die Beobachtung, die bereits bei der Optimierung der vorgeschlagenen Steuerungsfunktion in Abb. 5.3 (a) gemacht wurde, dass im Bereich der positiven SNR$_{\text{IN}}$-Werte gemessen in Dezibel die Verbesserung der Sprachsignalverständlichkeit mit steigenden SNR$_{\text{IN}}$-Werten abnimmt. Sonst erreicht $\alpha^{\text{BS-opt}}$ eine Verbesserung von nur $0.4\,\%$ des absoluten STOI-Maßes gegenüber dem durchschnittlichen eingangsseitigen STOI$_{\text{IN}}$-Wert von $86.4\,\%$, was zugegebenermaßen nicht besonders viel ist und sicherlich der hoch nichtstationären Natur von Störsignalen der CHiME-3-Daten geschuldet ist.

Um den Zusammenhang zwischen der Verständlichkeit und der Qualität der prozessierten Sprachsignale gemessen jeweils in STOI- und MOS-LQO$_{\text{WB}}$-Werten genauer zu untersuchen, werden neben den Steuerungsfunktionen α^{OS}, $\alpha^{\text{OS-opt}}$ und $\alpha^{\text{BS-opt}}$, die am Ende vom



(a) STOI-Verbesserung bei α^{OS} und $\alpha^{\text{BS-opt}}$

(b) STOI über MOS-LQO$_{\text{WB}}$

Abbildung 5.6.: Sprachsignalverständlichkeit gemessen in STOI und die Sprachsignalqualität gemessen in MOS-LQO$_{\text{WB}}$ für verschiedene Steuerungsfunktionen auf den Daten der CHiME-3-Datenbank.

Abschnitt 5.3 definiert wurden, noch vier weitere Steuerungsfunktionen eingeführt:

4. α^{opt}: die im MSE-Sinn optimale Steuerungsfunktion (5.5),
5. $\alpha^{\text{BS-MAP}}$: die MAP-basierte Steuerungsfunktion (5.21) mit $\Delta\nu = 2$,
6. $\alpha^{\text{BS-0}}$: die Bayes-motivierte Steuerungsfunktion (5.21) mit $\Delta\nu = 0$,
7. $\alpha^{\text{BS-MMSE}}$: die MMSE-basierte Steuerungsfunktion (5.21) mit $\Delta\nu =$ -2.

Die durchschnittlichen MOS-LQO$_{\text{WB}}$- und STOI-Werte, die von allen sieben Steuerungsfunktionen erreicht werden, sind neben den durchschnittlichen MOS-LQO$_{\text{WB}}$- und STOI-Werten gemittelt über alle gestörten Sprachsignale am Systemeingang in Abb. 5.6 (b) dargestellt[5]. Von den drei im MSE-Sinn optimalen Steuerungsfunktionen α^{OS}, $\alpha^{\text{OS-opt}}$ und α^{opt} erreicht die konventionelle Steuerungsfunktion des MS-Verfahrens α^{OS} zwar die beste durchschnittliche Sprachsignalqualität, allerdings geschieht dies auf Kosten von Sprachsignalverständlichkeit, die hier unter den allen betrachteten Steuerungsfunktionen am schlechtesten ausfällt. Die beste Sprachsignalverständlichkeit liefert hier die $\alpha^{\text{OS-opt}}$, die jedoch im Gegenzug zu den Signalen mit schlechterer Qualität der Sprachsignale führt als dies bei α^{OS} der Fall ist. Die im MSE-Sinn optimale Steuerungsfunktion α^{opt} erreicht in etwa eine ähnliche Leistungsfähigkeit wie die $\alpha^{\text{OS-opt}}$ Steuerungsfunktion mit leicht besseren MOS-LQO$_{\text{WB}}$- und schlechteren STOI-Werten. Somit scheint die Steuerungsfunktion α^{opt}, die durch die MSE-Minimierung (5.4) entsteht, eher zur besseren Sprachsignalverständlichkeit zu führen als zur besseren Qualität der prozessierten Sprachsignale. Insgesamt bestätigen diese Ergebnisse die These, dass die Qualität und die Verständlichkeit der Sprachsignale zwei konkurrierende Ziele sind. Ansonsten bleibt der Eindruck, dass die Verbesserung der Qualität der entstörten Sprachsignalen mit den konventionellen Systemen zur spektralen Sprachsignalentstörung leichter zu erreichen ist als die Verbesserung der Sprachsignalverständlichkeit.

Die Experimente mit den vier Bayes-motivierten Steuerungsfunktionen $\alpha^{\text{BS-opt}}$, $\alpha^{\text{BS-MAP}}$, $\alpha^{\text{BS-0}}$ und $\alpha^{\text{BS-MMSE}}$ offenbaren einen ähnlichen Zusammenhang zwischen den MOS-LQO$_{\text{WB}}$- und STOI-Werten. Bemerkenswert ist hier, dass die Steuerungsfunktion $\alpha^{\text{BS-opt}}$ auch auf den CHiME-3-Daten die beste Sprachsignalverständlichkeit unter allen Steuerungsfunktionen erreicht. Für die Verbesserung der Qualität der Sprachsignale soll hier die Steuerungsfunktion $\alpha^{\text{BS-MAP}}$ gewählt werden, die allerdings gegenüber der Funktion α^{OS} in der Sprachsignalqualität leicht verliert.

5.5. Zusammenfassung

Dieses Kapitel beschäftigte sich mit dem *Minimum Statistics* Verfahren aus [Mar01], das in den Systemen zur spektralen Sprachsignalentstörung für RLDS-Schätzung sehr häufig eingesetzt wird. Speziell wurde hier die verwendete Steuerungsfunktion des zeitvarianten Glättungsfaktors kritisch untersucht, welcher im MS-Verfahren in einer rekursiven Glättung momentaner spektraler Leistungen eines gestörten Sprachsignals verwendet wird. Obwohl diese Steuerungsfunktion im MSE-Sinne optimal gewählt wird, stellte sich immer noch die

[5] Man beachte, um die Analyse der experimentellen Ergebnisse zu erleichtern, sind in Abb. 5.6 (b) die Steuerungsfunktionen, die auf der MSE-Optimierung (5.5) basieren, durch blaue Diamanten und die Steuerungsfunktionen, die der Bayes-motivierten Optimierung entstammen, durch rote Kreise dargestellt.

Frage, ob sie auch hinsichtlich solche Bewertungsmaße wie Sprachsignalqualität und Verständlichkeit die bestmögliche Wahl ist oder ob es diesbezüglich bessere Alternativen gibt. Nachdem die Vor- und Nachteile der konventionellen Steuerungsfunktion diskutiert wurden, wurde eine alternative Steuerungsfunktion vorgeschlagen. Diese beruht auf einer Bayesschen Schätzung der spektralen Rauschleistungsdichte, die im Unterschied zum MS-Verfahren nicht als Parameter sondern als Zufallsvariable modelliert wird, die einer skalierten inversen Chi-Quadrat-Verteilung mit einem Freiheitsgrad-Parameter ν unterliegt, der mit Hilfe einer empirisch gewählten Funktion gesteuert wird. Die neue Steuerungsfunktion enthält dabei einen frei wählbaren Parameter $\Delta\nu$, der im Rahmen einer experimentellen Optimierung auf den gestörten Sprachdaten zu $\Delta\nu = -0.3$ so festgelegt wird, dass er im Mittel zu den entstörten Signalen mit bester Sprachsignalverständlichkeit gemessen mit dem STOI-Maß führt.

Laut den durchgeführten Untersuchungen verhinderte die Verwendung der neuen Steuerungsfunktion mangelnde Verfolgungsfähigkeit der rekursiven Glättung mit der konventionellen Steuerungsfunktion besonders in den Zeit-Frequenz-Bereichen, die vom Störsignal dominiert werden. In einer experimentellen Untersuchung auf den TIMIT-Daten, die von den verschiedenen Störsignalen der SPIB-Datenbank gestört wurden, zeigte sich, dass die neue Steuerungsfunktion eine Überschätzung der RLDS-Referenz verhinderte und dadurch eine leichte Verbesserung der Sprachsignalverständlichkeit bewirkte. Allerdings erwies sich die konventionelle Steuerungsfunktion als die bessere Wahl hinsichtlich der Sprachsignalqualität wie mit dem MOS-LQO$_{\text{WB}}$-Maß gemessen. Auch in der finalen Untersuchung auf den CHiME-3-Daten führte die alternative Steuerungsfunktion zu den prozessierten Sprachsignalen mit bester Sprachsignalverständlichkeit. Außerdem zeigte sich, dass eine Erhöhung des Parameters auf $\Delta\nu = 2$ zu einer vergleichbaren Verbesserung der Sprachsignalqualität führt, die von der konventionellen Steuerungsfunktion erzielt wird. Jedoch muss man dabei Verluste in der Verständlichkeit hinnehmen. Somit lässt sich durch die Wahl von $\Delta\nu$ leicht regulieren, welche Art der Verbesserung durch spektrale Sprachsignalentstörung erbracht werden soll. Ferner ergab die Untersuchung, dass eine Anpassung des Parameters α_{min} der konventionellen Steuerungsfunktion zum ähnlichen Zusammenhang zwischen Qualität und Verständlichkeit der Signale führte. Dadurch wurde ersichtlich, dass eine gleichzeitige Verbesserung der beiden Maße im Rahmen der verwendeten spektralen Entstörung nicht erreicht werden kann.

6. RLDS-Schätzung unter Verwendung eines neuronalen Netzes

Wie in Abschnitt 4.1 erwähnt, werden neuronale Netze aufgrund ihrer herausragenden Leistungsfähigkeit in der modernen digitalen Sprachsignalverarbeitung zunehmend eingesetzt. Einige Beispiele zu ihrem Einsatz in der einkanaligen spektralen Sprachsignalentstörung sind in Abschnitt 3.5 zu finden, wo die Netze entweder die komplette Entstörungsaufgabe mittels einer Regression übernehmen oder einen der Systembausteine aus Abb. 2.1 ersetzen. Im letzten Fall bilden die modellbasierten Verfahren in Kombination mit einem oder mehreren neuronalen Netzen ein Hybridsystem, das zur besseren Generalisierbarkeit, zur robusteren Schätzung und somit zur höheren Leistungsfähigkeit hinsichtlich der Sprachsignalentstörung führt [MT16]. Diese Beobachtung dient als Motivation, in diesem Kapitel ein neues Hybridsystem zu entwickeln, in dem ein neuronales Netz die Schätzung der Rauschleistungsdichte mit einer robusten NPP-Schätzung unterstützt. Also ist die Hauptaufgabe des neuronalen Netzes, im Spektrogramm des gestörten Sprachsignals die Zeit-Frequenz-Punkte zu finden, die vom Störsignal dominiert werden. Dabei muss die NPP-Schätzung kausal sein und somit nur auf den Daten der vergangenen STFT-Rahmen arbeiten. Zweifellos ist ein NPP-Schätzer mit einem VAD- oder einem SPP-Schätzer eng verwandt, die bei RLDS-Schätzung wie in [Hir93, HE95, GH11, GH12] oft verwendet werden und in der akustischen Quellentrennung häufig auch als spektrale Maskenschätzer bezeichnet werden. Dennoch wird der DNN-basierte Schätzer hier als ein NPP-Schätzer bezeichnet und bekommt somit einen Eigennamen, der seine Zielsetzung verdeutlicht.

Da der RLDS-Schätzer, der in diesem Kapitel entwickelt wird, von einem DNN-basierten spektralen Maskenschätzer Gebrauch macht, wird in Abschnitt 6.1 auf einige moderne Maskenschätzer kurz eingegangen, in denen neuronale Netze zum Einsatz kommen. Anschließend wird in Abschnitt 6.2 der DNN-basierte NPP-Schätzer ausführlich vorgestellt, wobei ein besonderer Schwerpunkt auf die Netzarchitektur gelegt wird. Um einen fairen experimentellen Vergleich des vorgeschlagenen RLDS-Schätzers mit den konventionellen Schätzern aus Unterabschnitt 3.1.1 zu ermöglichen, werden die letzten auf der Datenbank optimiert, die beim Training des neuronalen Netzes zum Einsatz kam. Eine ausführliche Beschreibung dieser Parameteroptimierung ist in Abschnitt 6.3 zu finden. Die Ergebnisse einer umfangreichen experimentellen Untersuchung, die eine klare Überlegenheit des vorgeschlagenen RLDS-Schätzers aufzeigt, werden zum Schluss in Abschnitt 6.4 erläutert. Dabei wird festgestellt, dass der DNN-basierte RLDS-Schätzer nicht nur den kleinsten Schätzfehler und die niedrigste Schätzfehlervarianz unter allen betrachteten RLDS-Schätzern aufweist, sondern auch zu den entstörten Sprachsignalen mit bester Qualität und gleichzeitig mit stärkster Störsignalunterdrückung führt. Man beachte, dass der in diesem Kapitel zu entwickelnde DNN-basierte RLDS-Schätzer zum ersten Mal in [CHDHU16] vorgestellt wurde.

6.1. Neuronale Netze zur Schätzung von spektralen Masken

In verschiedenen Systemen, die auf eine VAD- oder eine SPP-Schätzung angewiesen sind, werden statt modellbasierter Verfahren künstliche neuronale Netze eingesetzt. In manchen solchen Systemen wird allerdings statt einer SPP eine *ideal ratio mask* (IRM) geschätzt, die ebenfalls als eine SPP interpretiert werden kann. Die Zielsetzung eines SPP- und eines IRM-Schätzers ist in der Regel eine ideale binäre Maske (IBM), die von DeLiang Wang zum ultimativen Ziel eines Systems zur Trennung eines Sprachsignals von einer beliebigen Störquelle erklärt wird [Wan05].

Merkmalsentstörung mit einer DNN-basierten VAD-Schätzung: So wird in [ZW13] eine *denoising* DNN (DDNN) basierte VAD-Schätzung vorgestellt, die in zwei Schritten realisiert wird. Im ersten Schritt wird ein DDNN zur Merkmalsentstörung aufgebaut, das schichtweise wie ein konventioneller *denoising* Autoencoder unüberwacht vortrainiert wird. Als Eingangsdaten dienen dabei verschiedene Merkmale wie DFT-, MFCC- und LPC-Koeffizienten, die zu einem langen Vektor mit insgesamt 273 Komponenten zusammengesetzt werden. Das verwendete DDNN ist relativ klein und besteht aus drei verborgenen Feed-Forward-Schichten mit jeweils 54, 7 und 7 Knoten. Im zweiten Schritt wird das vortrainierte DDNN um eine zusätzliche Ausgangsschicht erweitert, welche die angestrebte VAD-Entscheidung für jeden Rahmen realisiert. Das erweiterte neuronale Netz wird anschließend als ein linearer Klassifikator überwacht nachtrainiert, bevor es als ein DDNN-basierter VAD-Klassifikator verwendet wird. In einer experimentellen Untersuchung auf den Daten der AURORA-2-Datenbank zeigte der netzwerkbasierte VAD-Schätzer eine gute Leistungsfähigkeit in einer ASR-Aufgabe eingesetzt in unterschiedlichen Störumgebungen.

Merkmalsentstörung mit einer zweistufigen DNN-basierten IRM-Schätzung: Sowohl die IRM-Schätzung als auch die Merkmalsentstörung in [NW13] finden im MSK-Bereich statt. Dabei wird in der ersten Systemstufe in jedem Frequenzband ein *feed-forward* DNN eingesetzt, das ein momentanes SNR in diesem Frequenzband schätzt, das mit Hilfe einer vordefinierten Sigmoid-Funktion in die IRM-Schätzwerte umgerechnet wird. An den Eingang eines Netzes werden dabei 103 verschiedene Merkmalen angelegt, die aus den Daten des aktuellen Rahmens berechnet werden. Außerdem hat das Netz zwei verborgene Schichten mit jeweils 200 Knoten und einen einzigen Knoten in der Ausgangsschicht. Die von der ersten Stufe geschätzten IRM-Werte nutzen allerdings keine Informationen aus den benachbarten Rahmen oder Frequenzbändern. Aus diesem Grund wird noch ein zweites DNN aufgebaut, dessen Eingang die Schätzwerte angelegt werden, die von den Netzen der ersten Systemstufe berechnet werden. Das zweite Netz realisiert eine zweidimensionale Glättung in der Zeit-Frequenz Ebene und glättet dabei über 9 benachbarte Frequenzbänder und 11 aufeinander folgenden Rahmen. In den durchgeführten ASR-Experimenten erzielt der vorgeschlagene IRM-Schätzer die kleinsten Wortfehlerraten.

Akustisches Beamforming mit einer BLSTM-basierten SPP-Schätzung: Die spektralen Masken müssen allerdings nicht unbedingt direkt auf die gestörten Merkmale multiplikativ angewandt werden, sondern könne auch für Schätzung der Systemparameter verwendet werden. So kann die Verwendung einer robusten DNN-basierten SPP-Schätzung in einem modellbasierten Verfahren zur Steigerung der Leistungsfähigkeit führen [MT16]. Ein geeignetes Beispiel aus dem Bereich der mehrkanaligen akustischen Strahlenformung wird in [HDCHU15] vorgestellt. Hier werden von einem DNN zwei unterschiedliche spektrale Masken für eine NPP und eine SPP berechnet, die für die Berechnung von Kreuzko-

varianzmatrizen des Störsignals und des Sprachsignals verwendet werden. Aus den letzten wird ein *generalized eigenvalue* (GEV) Strahlformer bestimmt, mit dem eine spektrale Filterung von mehrkanaligen gestörten Sprachsignalaufnahmen der CHiME-3-Datenbank durchgeführt wird. Bevor das prozessierte Signal an das Back-End weitergegeben wird, wird das einkanalige Ausgangssignal des GEV-Strahlformers noch in einem *blind analytical normalization* (BAN) Postfilter verarbeitet, der für eine möglichst verzerrungsfreie Signalverarbeitung sorgt [WHU07].

Ein DNN-basierter NPP/SPP-Schätzer besteht hier aus vier verborgenen Schichten - einer bidirektionalen LSTM (BLSTM) Schicht und drei Feed-Forward Schichten mit jeweils 256 (für jede der beiden Richtungen), 513, 513 und 1026 Knoten. An den Eingang des Netzes werden rahmenweise die spektralen Amplituden der 513 STFT-Koeffizienten angelegt, die aus einem gestörten Sprachsignal, aufgenommen von einem der sechs Mikrofone, berechnet werden. Als Kostenfunktion verwendet man die Kreuzentropie, wobei der Ausgang des Netzes mit einer IBM verglichen wird, die während des Trainings aus den STFT-Amplituden des ungestörten Sprachsignals und des Störsignals energiebasiert bestimmt wird. Im Unterschied zu den modellbasierten Schätzern, welche bei der SPP-Schätzung entweder die einzelnen Frequenzbände als statistisch unabhängig voneinander betrachten oder nur die Korrelationen zwischen den unmittelbar benachbarten Frequenzbändern berücksichtigen, erfasst der DNN-basierte SPP-Schätzer hier die Korrelationseigenschaften eines Sprachsignals entlang der gesamten Frequenzachse. Somit werden in der Testphase für jedes der sechs Mikrofone insgesamt sechs NPP- und sechs SPP-Masken berechnet, die mit Hilfe einer elementweisen Median-Operation zu den finalen NPP- und SPP-Masken zusammengeführt werden. In der CHiME-3-Aufgabe lieferte dieses System die besten ASR-Ergebnisse unter den Systemen, die nur zwei der insgesamt sieben Komponenten des Baseline-Systems veränderten [BMVW16].

6.2. Kausale DNN-basierte RLDS-Schätzung

Inspiriert von der bestechenden Einfachheit der VAD- oder SPP-basierten RLDS-Schätzer aus [Hir93, HE95, GH11, GH12], die laut Tab. 3.1 neben einer VAD- oder SPP-Schätzung nur eine ausgangsseitige Glättung benötigen, lässt sich ein RLDS-Schätzwert mit Hilfe einer einfachen Rekursionsgleichung erster Ordnung berechnen

$$\hat{\lambda}_D(k,\ell) = (1 - M_D(k,\ell)) \cdot \hat{\lambda}_D(k,\ell-1) + M_D(k,\ell) \cdot |Y(k,\ell)|^2, \tag{6.1}$$

wenn ein robuster NPP-Schätzer $M_D(k,\ell)$ gegeben ist, der auch als spektrale Maske für das Rauschsignal bezeichnet werden kann. Die spektrale Maske $M_D(k,\ell)$ stellt sich dabei als ein zeitvarianter Glättungsparameter dar, von dem der RLDS-Schätzer gesteuert wird. Die Gleichung (6.1) ist eng verwandt mit der Gleichung (22) in [GH12], in der allerdings eine SPP-Schätzung die Rolle der treibenden Kraft übernimmt. Im Unterschied zum Schätzverfahren aus [GH12], das aufgrund einer verhältnismäßig hohen Schätzfehlervarianz des dort verwendeten SPP-Schätzers eine zusätzliche ausgangsseitige Glättung der RLDS-Schätzwerte benötigt, wird im vorgeschlagenen Schätzer keine zusätzliche Glättung vorgesehen. Der Grund dafür sind die herausragenden Eigenschaften der DNN-basierten NPP-Schätzung, auf die im weiterem näher eingegangen wird. Der Verzicht auf die ausgangsseitige Glättung entspricht der Verwendung des Glättungsparameters $\alpha_{\text{pow}} = 0$ in [GH12].

Schicht	Neuronen	Typ	Nichtlinearität	$\mathcal{P}_{\text{dropout}}$
L1	512	LSTM	Tanh	0.5
L2	1024	FF	ELU	0.5
L3	1024	FF	ELU	0.5
L4	513	FF	Sigmoid	0

Tabelle 6.1.: Architektur des verwendeten neuronalen Netzes für eine NPP-Schätzung bei FFT Länge $K = 2^{10}$.

DNN-Architektur: Für die Schätzung der spektralen Maske $M_D(k,\ell)$ wird das DNN aus [HDCHU15] verwendet, das für die kausale Signalverarbeitung angepasst wird. Die resultierende Netzarchitektur ist in Tab. 6.1 für die FFT-Länge $K = 2^{10}$ zusammengefasst. Das verwendete neuronale Netz beinhaltet drei verborgene Schichten L1-L3 und eine Ausgangsschicht L4. An den Eingang des Netzes werden die spektralen Amplituden aller STFT-Koeffizienten des aktuellen ℓ-ten Rahmes $[|Y(0,\ell)|, |Y(1,\ell)|, \ldots, |Y(\frac{K}{2}),\ell|]$ angelegt. Da im Unterschied zu [HDCHU15] für die RLDS-Schätzung kein SPP-Schätzer benötigt wird, kann die Größe der Ausgangsschicht halbiert werden. Aus den spektralen Amplituden am Eingang berechnet das Netz für jeden ℓ-ten Rahmen die Werte der spektralen Maske $[|M_D(0,\ell)|, |M_D(1,\ell)|, \ldots, |M_D(\frac{K}{2}),\ell|]$ Die Forderung einer kausalen Signalverarbeitung verlangt zwei weitere Anpassungen der Netzarchitektur aus [HDCHU15], die zum einen die erste verborgene Schicht und zum anderen die Normalisierungsart des Datenflusses im Netz betreffen. Im Rahmen der ersten Anpassung wird eine BLSTM-Schicht durch eine einfachere LSTM-Schicht ersetzt, die als erste verborgene Schicht L1 zum Einsatz kommt. Da die LSTM-Schicht sich als eine spezielle Art eines rekurrenten Netzes darstellt, ermöglicht sie die Erfassung der zeitlichen Korrelationen von Sprachsignalen. Um die Abwesenheit des BLSTM-Rückwärtspfades zu kompensieren, wird die Anzahl der Knoten in der LSTM-Schicht verdoppelt und auf 512 festgelegt. Zusätzlich wird auch die Knotenanzahl der beiden darauf folgenden verborgenen Feed-Forward-Schichten auf 2^{10} erhöht. Im Rahmen der zweiten Anpassung muss die Batch-Normalisierung des Datenflusses im Netz aus [IS15] vermieden werden, die für eine kausale Signalverarbeitung in der Testphase ungeeignet ist. Die Berechnung der Normalisierungsstatistiken für die Testdaten wird dadurch vermieden, dass die Testdaten mit den Statistiken normalisiert werden, die während des Trainings gelernt werden. Sonst wird in den Schichten L2 und L3 statt der *rectified linear unit* (ReLU) Funktion die *exponential linear unit* (ELU) Funktion als Nichtlinearität verwendet [CUH15].

Kostenfunktion und Zielmaske: Als Kostenfunktion wird die binäre Kreuzentropie (engl. *binary cross entropy*, BCE) zwischen einer idealen binären Maske des Störsignals $\text{IBM}_D(k,\ell)$ und dem Ausgang des Netzes $M_D(k,\ell)$ verwendet, die folgendermaßen berechnet wird:

$$\begin{aligned}\text{BCE} = -\frac{1}{K\cdot L}\sum_{k=1}^{K}\sum_{\ell=1}^{L}\{&\text{IBM}_D(k,\ell)\cdot\log_2 M_D(k,\ell)\\ &+(1-\text{IBM}_D(k,\ell))\cdot\log_2(1-M_D(k,\ell))\}.\end{aligned} \tag{6.2}$$

Die Zielmaske $\text{IBM}_D(k,\ell)$ wird dafür aus den spektralen Amplituden des Störsignals und

des ungestörten Sprachsignals, die in der Trainingsphase vorliegen, wie folgt ausgerechnet

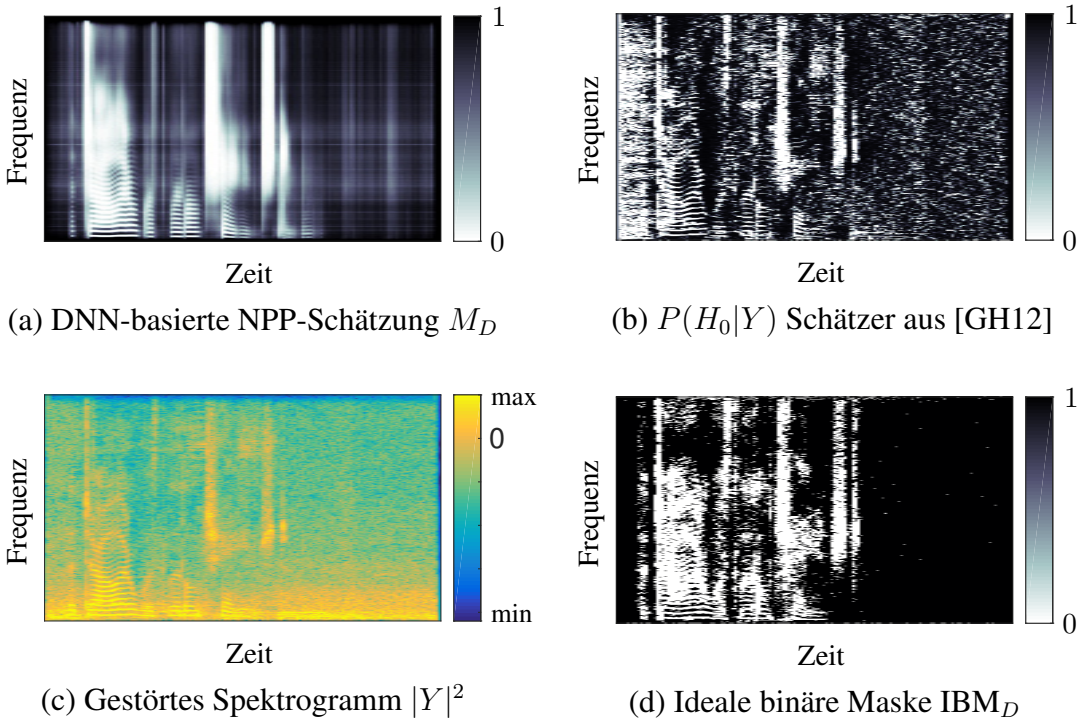
$$\text{IBM}_D(k,\ell) = \begin{cases} 1, & \frac{|D(k,\ell)|}{|S(k,\ell)|} > \vartheta_D \\ 0, & \text{sonst} \end{cases} \tag{6.3}$$

mit der Schwelle $\vartheta_D = 10$. Somit wird ein Zeit-Frequenz-Punkt nur dann dem Störsignal zugeschrieben, wenn dieser vom Störsignal sehr stark dominiert wird. Auf diese Weise wird die Zuordnung der Zeit-Frequenz-Punkte mit einer schwachen spektralen Sprachsignalenergie zur spektralen Maske des Störsignals vermieden. Die Verwendung solcher konservativen Zielmasken während des Trainings des Netzes führt in der Testphase zu einer robusten RLDS-Schätzung mit einer reduzierten Schätzfehlervarianz, wie im Weiteren gezeigt wird. Man beachte, dass die resultierende spektrale Maske reellwertig im Wertebereich $M_D(k,\ell) \in (0\,;1)$ ist, auch wenn die binären idealen Maske $\text{IBM}_D(k,\ell)$ als Zielmaske verwendet wird, die nur die Werte 0 oder 1 annimmt. Somit kann $M_D(k,\ell)$ als Wahrscheinlichkeit interpretiert werden, die in (6.2) einfließt und aus diesem Grund als NPP-Schätzung bezeichnet wird.

DNN-Training: Der DNN-basierte NPP-Schätzer aus Tab. 6.1 wird auf den Trainingsdaten der CHiME-3-Datenbank trainiert, siehe Abschnitt 2.4. Dafür werden die mit $16\,\text{kHz}$ abgetasteten Audiosignale einer STFT unterzogen mit der FFT-Länge von $K = 2^{10}$, dem Rahmenvorschub $R = 2^8$ und dem Blackman-Analysefenster wie in [HDCHU15]. Da ein erfolgreiches Training eines neuronalen Netzes mit einer großen Menge an Trainigsdaten einhergeht, werden für das Training die Audioaufnahmen aller sechs Mikrofone verwendet [IS15]. Die Initialwerte von Gewichten aller Schichten werden aus einer Gleichverteilung $\omega_0 \sim \mathcal{U}\left[-a_\omega, a_\omega\right]$ gezogen, wobei a_ω ein fester Parameter ist. Während für die LSTM-Schicht $a_\omega = 0.04$ gewählt wird, ist für die anderen Schichten $a_\omega = \sqrt{6/(N_{\text{IN}} + N_{\text{OUT}})}$ gesetzt, wobei N_{IN} und N_{OUT} jeweils die Anzahl der Eingänge einer Schicht (oder der Knoten der vorigen Schicht) und die Anzahl der Ausgänge einer Schicht (oder der eigenen Knoten) sind [GB10]. Um bessere Generalisierung zu erreichen, wird in den ersten drei Schichten die *Dropout*-Technik mit den Wahrscheinlichkeiten $\mathcal{P}_{\text{dropout}} = 0.5$ verwendet [SHK+14]. Während bei der LSTM-Schicht nur die Eingang-zum-Knoten-Verbindungen von der Dropout-Technik betroffen werden [ZSV14], erfährt die letzte Schicht des Netzes keine derartige Regularisierung. Das Training des Netzes wird nach 8 Trainingsepochen abgeschlossen. Weitere Einzelheiten zum Training des Netzes sind in [CHDHU16] zu finden[1].

Um zu verdeutlichen, wie die spektralen Masken $M_D(k,\ell)$ aussehen, die der vorgeschlagene DNN-basierte NPP-Schätzer berechnet, wird in Abb. 6.1 (a) eine exemplarische Maske vorgestellt, die auf einer Audioaufnahme der CHiME-3-Datenbank aus dem *isolated simulated development* Datensatz mit dem Namen *f04_053c010p_str* geschätzt wird. Diese kurze Audioaufnahme der gesamten Dauer von etwa 2.3 Sekunden wurde auf einer Straße aufgenommen und weist das eingangsseitige SNR von etwa $5\,\text{dB}$ auf. Zum Vergleich wird in Abb. 6.1 (b) auch die entsprechende NPP-Schätzung $P(H_0|Y) = 1 - P(H_1|Y)$ abgebildet, die aus den SPP-Schätzwerten $P(H_1|Y)$ des RLDS-Schätzers aus [GH12] berechnet wird. Die beiden Masken werden aus dem Spektrogramm des gestörten Sprachsignals geschätzt, das in Abb. 6.1 (c) dargestellt ist. Außerdem wird in Abb. 6.1 (d) die dazugehörige ideale binäre Maske $\text{IBM}_D(k,\ell)$ für die Schwelle $\vartheta_D = 1$ präsentiert. Wie man sieht, ist

[1] Besonderer Dank gilt an dieser Stelle Jahn Heymann und Lukas Drude sowohl für softwaretechnische Realisierung und Training des verwendeten Netzes als auch für Berechnung DNN-basierter Masken.



(a) DNN-basierte NPP-Schätzung M_D

(b) $P(H_0|Y)$ Schätzer aus [GH12]

(c) Gestörtes Spektrogramm $|Y|^2$

(d) Ideale binäre Maske IBM_D

Abbildung 6.1.: Beispiel einer DNN-basierten NPP-Schätzung $M_D(k, \ell)$ auf einer Äußerung der CHiME-3-Datenbank mit dem Namen *f04_053c010p_str* im Vergleich zur entsprechenden NPP-Schätzung mit dem Verfahren aus [GH12].

die spektrale Maske des DNN-basierten NPP-Schätzers ziemlich geglättet, was auch das Hauptargument für den Verzicht auf die zusätzliche Ausgangsglättung ist. Dabei ist das neuronale Netz durchaus fähig, die feinen Strukturen eines Sprachsignals entlang der Zeitachse ziemlich genau zu rekonstruieren. Da beim DNN-basierten Schätzer in der Testphase angenommen wird, dass die Frequenzbänder $k < 5$ und $k > 500$, die den Frequenzen unterhalb von $\sim 78\,\text{Hz}$ und oberhalb von $\sim 7.8\,\text{kHz}$ entsprechen, keine signifikanten Sprachsignalanteile beinhalten, werden die entsprechenden Werte der spektralen Maske zu $M_D(k, \ell) = 1$ gesetzt [SA05]. Ein weiterer Vorteil des DNN-basierten Schätzers wird bei Betrachtung von breitbandigen Phonemen ersichtlich, die von ihrer Natur aus beinahe alle Frequenzen gleichzeitig belegen. Dadurch, dass an die Eingangsschicht des Netzes die spektralen Amplituden aller Frequenzbände gemeinsam angelegt werden, lernt das neuronale Netz die Strukturen eines Sprachsignals, die sich entlang der Frequenzachse erstrecken. Im Gegensatz dazu offenbart der Schätzer aus [GH12] die Nachteile wie Notwendigkeit einer Anlaufphase, ungeglättete Maskenstruktur und Streuen der Schätzwerte in den Pausen. Allerdings berechnet auch der DNN-basierte NPP-Schätzer in den Pausen eines Sprachsignals eine spektrale Maske, die mit einer IBM nicht ganz übereinstimmt.

Im Weiteren soll die Leistungsfähigkeit des DNN-basierten RLDS-Schätzers bei seinem Einsatz in einem System zur spektralen Sprachsignalentstörung auf dem *development* Datensatz der CHiME-3-Datenbank untersucht werden, der alle für eine Auswertung notwendigen Signale beinhaltet und in Abschnitt 2.4 bereits beschrieben wurde. Dabei soll der neu entwickelte Schätzer auch mit den anderen modernen RLDS-Schätzern verglichen werden, die in Unterabschnitt 3.1.1 beschrieben wurden. Außerdem soll der in Kap. 5 vorgestellte

RLDS-Schätzer mit der Bayes-motivierten Steuerungsfunktion in den Vergleich miteinbezogen werden. Allerdings befindet sich der vorgeschlagene RLDS-Schätzer gegenüber den konventionellen Schätzern eindeutig im Vorteil, denn sein DNN-basierter NPP-Schätzer wurde auf den Trainingsdaten der CHiME-3-Datenbank trainiert, sodass seine Parameter auf diese Daten ziemlich gut angepasst sind. Die Parameter der konventionellen Vergleichsverfahren zur RLDS-Schätzung wurden jedoch auf ganz anderen Daten festgelegt, die ihre Entwickler beim Entwurf der Schätzer verwendeten, und müssen nicht zwangsläufig für die CHiME-3-Datenbank optimal sein. Begründet in diesen Überlegungen sollen die Parameter der konventionellen RLDS-Schätzer zunächst auf den CHiME-3-Daten optimiert werden, damit ein fairer Vergleich mit dem DNN-basierten RLDS-Schätzer ermöglicht wird.

6.3. Optimierung der konventionellen Rauschschätzer

Die Festlegung eines Optimierungskriteriums ist für die bevorstehende Parameteroptimierung von entscheidender Bedeutung. Obwohl das Hauptziel eines RLDS-Schätzers eine möglichst präzise Schätzung der Rauschleistungsdichte ist, dient sein Schätzwert keinem Selbstzweck, sondern wird in erster Linie für die spektrale Sprachsignalentstörung eingesetzt. Aus diesem Grund sollen die zu optimierende Parameter so gewählt werden, dass sie auf der CHiME-3-Datenbank neben einer guten RLDS-Schätzung mit einem kleinen Schätzfehler und einer geringen Schätzfehlervarianz auch eine gute Qualität der entstörten Sprachsignale und eine starke Störsignalunterdrückung hervorrufen. So eine datengetriebene Parameteroptimierung kann nur nummerisch durchgeführt werden. Dabei sollen die Bewertungsmaße LEM, LEV, SNR_{OUT} und MOS-LQO gemeinsam verwendet werden, die in Abschnitt 2.3 eingeführt wurden. Dafür sollen sie zu einem kombinierten Bewertungsmaß zusammengeführt werden[2]. Da die erwähnten Bewertungsmaße von ihrer Größenordnung her verschiedene Wertebereiche belegen und unterschiedliche Zielrichtungen aufweisen[3], werden sie zunächst jeder für sich so auf den Wertebereich $[0;\, 1]$ normiert, dass eine gute Leistungsfähigkeit großen Werten des normierten Maßes entspricht und die schlechte Leistungsfähigkeit den kleinen. Das kombinierte Maß (KM) wird dann aus den normierten Bewertungsmaßen LEM_{norm}, LEV_{norm}, $\text{SNR}_{\text{OUT,norm}}$ und $\text{MOS-LQO}_{\text{norm}}$ wie folgt berechnet:

$$\text{KM} = \frac{\text{LEM}_{\text{norm}} + \text{LEV}_{\text{norm}} + \text{SNR}_{\text{OUT,norm}} + \text{MOS-LQO}_{\text{norm}}}{4}. \tag{6.4}$$

Man beachte, höhere Werte von KM gehen mit der besseren Leistungsfähigkeit eines RLDS-Schätzers einher. Da KM aus normierten Größen berechnet wird, liegt er im Bereich $[0,\, 1]$. Während LEM- und LEV-Maße direkt aus den RLDS-Schätzwerten $\hat{\lambda}_D(k, \ell)$ unter Verwendung des Spektrogramms des Störsignals $\lambda_D(k, \ell) = |D(k, \ell)|^2$ als RLDS-Referenz ausgerechnet werden, muss für die Berechnung der Maße SNR_{OUT} und MOS-LQO die spektrale Sprachsignalentstörung durchgeführt werden.

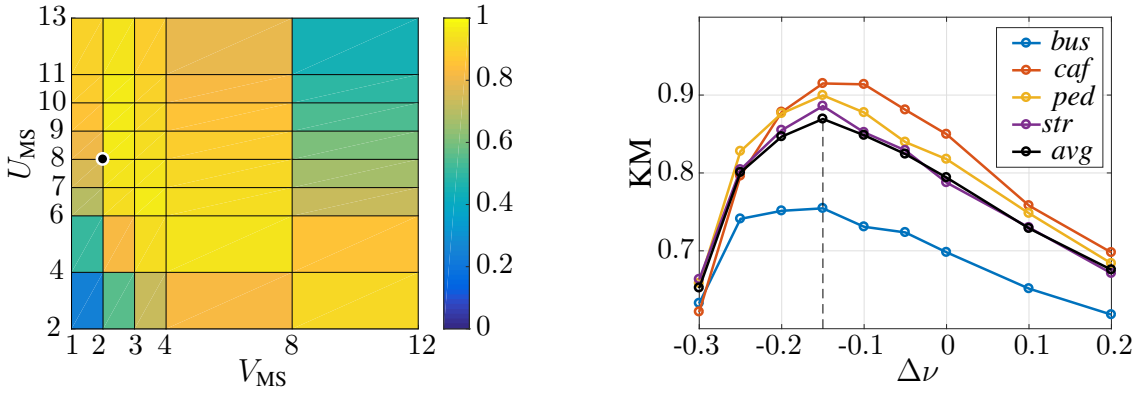
System zur spektralen Entstörung: Die spektrale Sprachsignalentstörung wird wie das System aus Abb. 2.2 mit dem einzigen Unterschied aufgebaut, dass auf den Baustein 4 ver-

[2] Verwendung von kombinierten Bewertungsmaßen ist eine gängige Praxis bei Optimierung der Schätzer. So werden in [YF11] drei verschiedene Messgrößen zu einem kombinierten *figure of merit* (FoM) Maß zusammengesetzt, um den Gewichtsfaktor des DD-Verfahrens zur *a priori* SNR-Schätzung zu optimieren.

[3] Während kleinere Werte von LEM und LEV angestrebt werden, sind höhere Werte von MOS-LQO und SNR_{OUT} wünschenswert.

zichtet wird. Als Baustein 1 werden zehn verschiedene RLDS-Schätzer eingesetzt, die optimiert werden sollen. Hinsichtlich der Implementierung der Rauschschätzer soll erwähnt werden, dass der Quellcode folgender fünf Schätzer entweder von den Entwicklern direkt zur Verfügung gestellt oder von den öffentlich zugänglichen Informationsquellen abgerufen wurde: OSMS aus [Mar01], MCRA aus [CB02], IMCRA aus [Coh03], MMSE-BM aus [HHJ10] und SPP-FP aus [GH12]. Die restlichen Vergleichsverfahren zur RLDS-Schätzung wurden selbstständig entsprechend den jeweiligen Veröffentlichungen implementiert. Die Daten des ersten Rahmens werden für die Initialisierung der Schätzer verwendet. Sonst wird es keine weitere Annahme hinsichtlich der Sprachsignalabwesenheit am Anfang einer gestörten Audioaufnahme gemacht. Im Baustein 2 wird der MMSE-LSA-Schätzer der spektralen Amplituden aus (2.27) als spektrale Filterfunktion $G_{\mathrm{LSA}}(k,\ell)$ benutzt [EM85], die über (2.19) von unten mit $G_{\min} = -18\,\mathrm{dB}$ begrenzt wird. Im Baustein 3 wird das DD-Verfahren mit dem Gewichtsfaktor $\alpha_{\mathrm{DD}} = 0.98$ in (3.13) als ein *a priori* SNR-Schätzer mit einer unteren Grenze von $\xi_{\min} = -18\,\mathrm{dB}$ verwendet [EM84]. Für die Parameteroptimierung werden $25\,\%$ des *isolated simulated development* Datensatzes verwendet, die im Weiteren als Optimierungsdaten bezeichnet werden. Diese bestehen aus insgesamt 410 kurzen Äußerungen, in denen alle vier vorhandenen Störumgebungen *bus*, *caf*, *ped* und *str* in etwa im gleichen Maße vertreten sind.

Parameteroptimierung am Beispiel des OSMS-Schätzers: Da einige RLDS-Schätzer eng miteinander verwandt sind und dadurch diverse Parameter mit einer ähnlichen Funktionalität besitzen, werden sie gemeinsam optimiert. So teilen der OSMS-Schätzer und der in Kap. 5 vorgestellte *Bayes-motivated smoothed* MS (BSMS) Schätzer zwei Parameter für die Realisierung der Minimumsuche, die sich in [Mar01] als die Anzahl der verwendeten Unterfenster U_{MS} und die Anzahl der Rahmen in einem Unterfenster V_{MS} darstellen, sodass die gesamte Fensterlänge für die Minimumsuche $D_{\mathrm{MS}} = U_{\mathrm{MS}} \cdot V_{\mathrm{MS}}$ beträgt. In diesem Fall ist es sinnvoll, zunächst den OSMS-Rauschschätzer hinsichtlich der Parameter U_{MS} und V_{MS} zu optimieren und anschließend mit den optimierten Parametern $U_{\mathrm{MS,opt}}$ und $V_{\mathrm{MS,opt}}$ den verbleibenden Parameter des BSMS-Verfahrens $\Delta\nu$ optimal zu wählen. Die gemeinsame Optimierung der OSMS- und BSMS-Schätzer wird in Abb. 6.2 verdeutlicht. In Abb. 6.2 (a) werden die optimalen Parameter $U_{\mathrm{MS,opt}} = 8$ und $V_{\mathrm{MS,opt}} = 2$ des OSMS-Rauschschätzers so gewählt, dass das KM-Maß, das hier über die dritte Dimension aufgetragen ist, maximiert wird. So-



(a) OSMS: $U_{\mathrm{MS,opt}} = 8$ und $V_{\mathrm{MS,opt}} = 2$

(b) BSMS: $\Delta\nu_{\mathrm{opt}} = -0.15$

Abbildung 6.2.: Optimierung von OSMS- und BSMS-Verfahren zur RLDS-Schätzung.

mit beträgt die gesamte Fensterlänge für die Minimumsuche nur $D_{\text{MS,opt}} = 16$ Rahmen, was für die gewählten STFT-Parameter einer Zeitdauer von etwa $0.25\,\text{s}$ entspricht. Die optimale Fensterlänge ist somit deutlich kleiner als die vom Autor des OSMS-Verfahrens empfohlene Fensterlänge, die bei den Parametern $U_{\text{MS,epm}} = 8$ und $V_{\text{MS,epm}} = 12$ mit einer Zeitdauer von etwa $1.5\,\text{s}$ einhergeht [Mar01]. Die empfohlene Wahl der Anzahl der Unterfenster $U_{\text{MS,opt}} = U_{\text{MS,epm}}$ wird aber durch Optimierung bestätigt. Die Reduzierung der frequenzunabhängigen Fensterlänge D_{MS} konnte damit erklärt werden, dass ein Sprachsignal bei tieferen und höheren Frequenzen unterschiedliche Strukturen im STFT-Bereich aufweist, siehe Abb. 6.1 (c). Während die tiefpasslastigen Formanten sich entlang der Zeitachse ausbreiten und somit längere Fensterlängen D_{MS} benötigten, erstrecken sich die breitbandigen Konsonanten entlang der Frequenzachse und besitzen schmale Strukturen entlang der Zeitachse, was zur Reduzierung der Fensterlänge D_{MS} beiträgt.

Die resultierende Verbesserung der einzelnen Bewertungsmaße, die durch Optimierung erreicht wird, wird in Tab. 6.2 festgehalten. Demnach verbessern sich alle Bewertungsmaße des OSMS-Schätzers auf den verwendeten Daten deutlich, was ein Zeichen der erfolgreichen Optimierung ist. Dabei ist zu beachten, dass auch der OSMS-Schätzer mit den empfohlenen Parametern sowohl zu einer leichten Verbesserung der Sprachsignalqualität als auch zu einer gewissen Störsignaldämpfung der stark gestörten Audioaufnahmen der CHiME-3-Datenbank führt. Die Reduzierung der Rahmenanzahl V_{MS} in einem Unterfenster von 12 auf 2 trägt jedoch zur besseren gesamten Leistungsfähigkeit das OSMS-Verfahrens bei, wofür eine Erhöhung des kombinierten Bewertungsmaßes von 0.37 auf beachtliche 0.96 spricht.

Parameteroptimierung des BSMS-Schätzers: Wie in Kap. 5 beschrieben, ist der BSMS-Schätzer eine Weiterentwicklung des OSMS-Verfahrens und realisiert dieselbe Minimumsuche. Aus diesem Grund ist es sinnvoll, die bereits optimierten Parameter $U_{\text{MS,opt}}$ und $V_{\text{MS,opt}}$ des OSMS-Verfahrens beim BSMS-Verfahren zu verwenden und nur den restlichen Parameter $\Delta\nu$ gesondert zu optimieren. Benutzt man die in Kap. 5 empfohlenen BSMS-Parameter $U_{\text{MS,emp}} = 8$, $V_{\text{MS,emp}} = 12$ und $\Delta\nu_{\text{emp}} = -0.3$ resultieren auf den Optimierungsdaten keine zufriedenstellende Ergebnisse, siehe die dritte Zeile von Tab. 6.2. Die Verwendung von $U_{\text{MS,opt}}$ und $V_{\text{MS,opt}}$ verbessert die Leistungsfähigkeit des BSMS-Verfahrens deutlich. Die Optimierung des Parameters $\Delta\nu$ für verschiedene Störumgebungen der CHiME-3-Datenbank ist in Abb. 6.2 (b) dargestellt, in der außerdem die über alle Störumgebungen gemittelte Kurve (*avg*) zu sehen ist. Der Verlauf der KM-Kurven offenbart die Notwendigkeit einer leichten

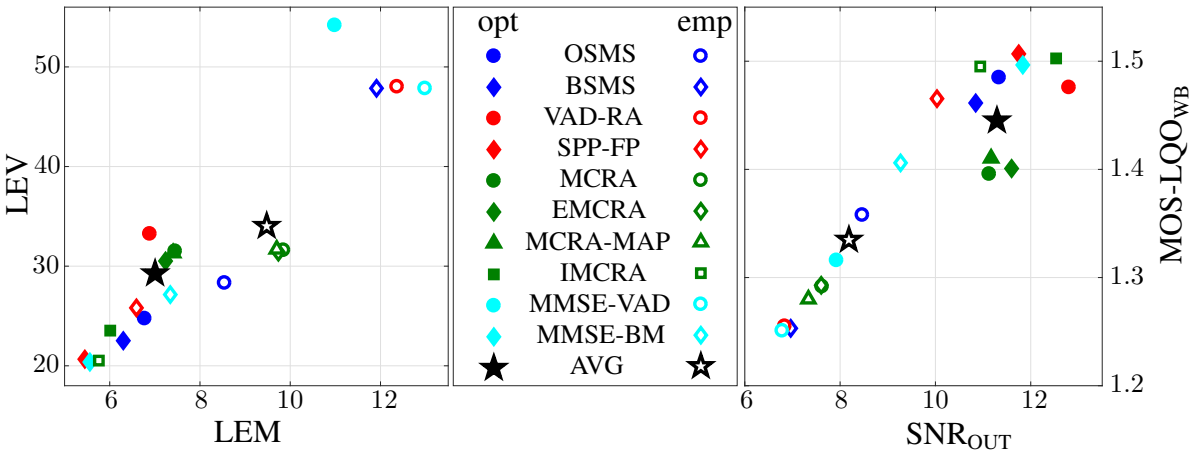
Schätzer	Parameter	LEM	LEV	$\text{SNR}_{\text{OUT}}/\text{dB}$	MOS-LQO	KM
OSMS	$V_{\text{MS,emp}} = 12$	8.54	28.33	8.47	1.36	0.37
	$V_{\text{MS,opt}} = 2$	6.77	24.75	11.3	1.48	0.96
BSMS	$V_{\text{MS,emp}}, \Delta\nu_{\text{emp}} = \text{-}0.3$	11.9	47.85	7.0	1.25	-
	$V_{\text{MS,opt}}, \Delta\nu_{\text{emp}}$	6.29	22.46	10.4	1.45	0.65
	$V_{\text{MS,opt}}, \Delta\nu_{\text{opt}} = \text{-}0.15$	6.30	22.52	10.8	1.46	0.87

Tabelle 6.2.: Leistungsfähigkeit der OSMS- und BSMS-Verfahren vor und nach Optimierung auf den Optimierungsdaten der CHiME-3-Datenbank mit den durchschnittlichen eingangsseitigen $\text{SNR}_{\text{IN}} = 6.1\,\text{dB}$ und $\text{MOS-LQO}_{\text{WB,IN}} = 1.21$.

Anpassung des empfohlenen Parameters $\Delta\nu_{\text{emp}}$, der ja auf den Daten der TIMIT-Datenbank bestimmt wurde. Gemittelt über alle Störumgebungen soll $\Delta\nu_{\text{opt}} = -0.15$ gewählt werden, was allerdings nur die Störsignaldämpfung leicht verbessert, wie die letzte Zeile von Tab. 6.2 zeigt. Dabei ist zu beachten, dass der optimale Parameter für eine relativ gute Verbesserung des kombinierten Maßes von 0.65 auf 0.87 sorgt. Sonst zeigt die Optimierung des BSMS-Verfahrens, dass es nicht immer möglich ist, alle Bewertungsmaße gleichzeitig zu verbessern. Außerdem ist aus Abb. 6.2 (b) deutlich, dass die gestörten Sprachsignale aufgenommen in einem Bus beim Einsatz eines BSMS-Rauschschätzers besonders schwer zu entstören sind. Verglichen mit dem optimierten OSMS-Verfahren erreicht der optimierte BSMS-Schätzer auf den Optimierungsdaten etwas bessere LEM- und LEV-Werte und etwas schlechtere SNR_{OUT}- und $\text{MOS-LQO}_{\text{WB}}$-Werte.

Optimierungsergebnisse: Ähnlich werden auch die anderen acht RLDS-Schätzer optimiert, dessen detaillierten Optimierungsergebnisse aus Platzgründen nicht weiter vorgestellt werden. Dafür werden die finalen Auswirkungen der Optimierung auf die Leistungsfähigkeit der RLDS-Schätzer in Abb. 6.3 vorgestellt, wo die Bewertungsmaße LEM, LEV, SNR_{OUT} und $\text{MOS-LQO}_{\text{WB}}$ abgebildet sind, die auf den Optimierungsdaten berechnet werden. Da die durchgeführte Optimierung nicht alle Bewertungsmaße gleichzeitig verbessern konnte, werden Ergebnisse der betrachteten RLDS-Schätzer sowohl mit den von den Entwicklern empfohlenen Parametern (emp) als auch mit den optimierten Parametern (opt) vorgestellt. Außerdem sind die durchschnittlichen Bewertungsmaße (AVG) dargestellt, die gemittelt über alle Verfahren zum einen mit den optimierten Parametern und zum anderen mit den empfohlenen Parametern berechnet werden. Die Ergebnisse der RLDS-Schätzer, die miteinander verwandt sind, bekommen die gleichen Darstellungsfarbe. So sind Bewertungsmaße der Schätzer OSMS und BSMS aus Tab. 6.2 in blau abgebildet.

Wie in Abb. 6.3 zu sehen ist, profitieren die beiden SPP- oder VAD-basierten RLDS-Schätzer von der datengetriebenen Optimierung sehr stark. Insbesondere verbessert das VAD-



(a) RLDS-Schätzung (b) Sprachsignalentstörung

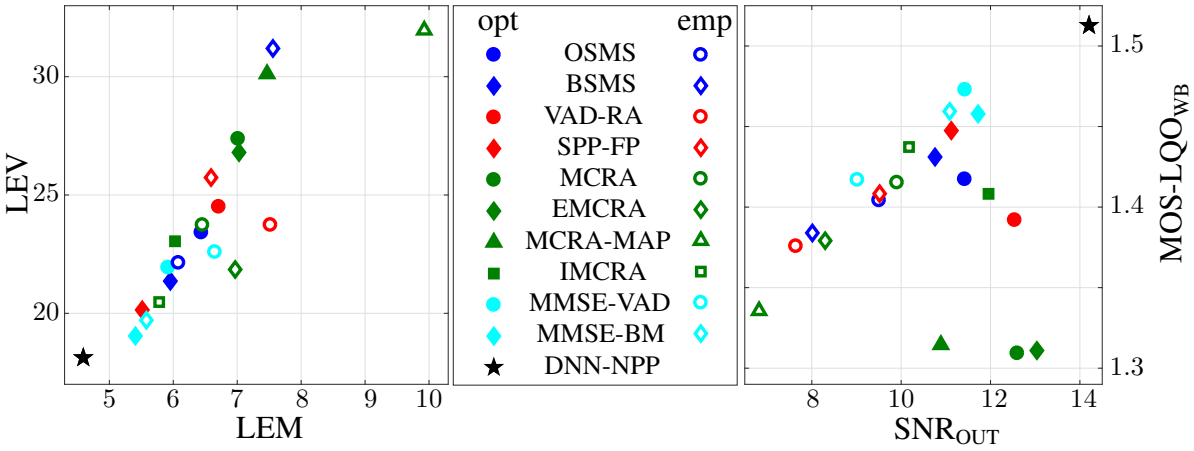
Abbildung 6.3.: Leistungsfähigkeit der zehn modernen RLDS-Schätzer für die optimierten Parameter (opt) und für die empfohlenen Parameter (emp): (a) Schätzfehler und Schätzfehlervarianz gemessen in LEM und LEV, (b) Störsignaldämpfung und Qualität der entstörten Sprachsignale gemessen in SNR_{OUT} und $\text{MOS-LQO}_{\text{WB}}$.

RA-Verfahren seine Leistungsfähigkeit in allen Bewertungsmaßen enorm, sodass er hinsichtlich der Störsignaldämpfung unter allen betrachteten RLDS-Schätzern am besten abschneidet. Das SPP-FP-Verfahren liefert dabei die beste Sprachsignalqualität gefolgt von MMSE-BM- und IMCRA-Verfahren. Das letzte liefert unter den vier MCRA-basierten Verfahren die besten Ergebnisse und profitiert von seiner ausgeklügelten Biaskorrektur und besseren Eigenschaften in der SPP-Schätzung bei schwacher spektraler Sprachsignalleistung [Coh03]. Die drei anderen MCRA-basierten Verfahren unterscheiden sich voneinander auf den Optimierungsdaten nicht all zu viel. Unter den MMSE-basierten Schätzern fällt eine starke Abhängigkeit des MMSE-VAD-Verfahrens von einer guten und zuverlässigen Initialisierung des Schätzers auf. Weiteres Nachforschen hat nachgewiesen, dass dieser Schätzer auf einige Rahmen ohne Sprachsignalpräsenz am Anfang einer gestörten Audioaufnahme angewiesen ist. Die durch AVG gekennzeichneten durchschnittlichen Werte der Bewertungsmaße verdeutlichen, dass eine datengetriebene Optimierung bei der Parametrisierung der RLDS-Schätzer durchaus empfehlenswert ist. So führte die durchgeführte Parameteroptimierung gemittelt über alle RLDS-Schätzer zur Reduzierung des Schätzfehlers gemessen in LEM von etwa 9.5 auf 7 und zur Verringerung der Schätzfehlervarianz gemessen in LEV von 34 auf 29.3, siehe Abb. 6.3 (a). Bessere RLDS-Schätzung steigerte die Leistungsfähigkeit der gesamten spektralen Entstörung. Während die RLDS-Schätzer mit den empfohlenen Parametern das globale eingangsseitige $\mathrm{SNR}_{\mathrm{IN}}$ von $6.1\,\mathrm{dB}$ auf nur $\mathrm{SNR}_{\mathrm{OUT}}$ von $8.2\,\mathrm{dB}$ verbessern, erreichen Systeme mit den optimal parametrisierten RLDS-Schätzern ein durchschnittliches ausgangsseitiges $\mathrm{SNR}_{\mathrm{OUT}}$ von $11.3\,\mathrm{dB}$, siehe Abb. 6.3 (b). Eine ähnliche Situation ist auch bei der Verbesserung der Sprachsignalqualität vorzufinden. Ausgehend von einem eingangsseitigen MOS-$\mathrm{LQO}_{\mathrm{WB,IN}}$-Wert von 1.21 liefert die empfohlene Parametrisierung das ausgangsseitige MOS-$\mathrm{LQO}_{\mathrm{WB,OUT}}$ von 1.33, während die RLDS-Schätzer mit den optimierten Parametern MOS-$\mathrm{LQO}_{\mathrm{WB,OUT}} = 1.45$ erreichen.

Nachdem die Parameter der modernen RLDS-Schätzer auf den CHiME-3-Optimierungsdaten optimiert sind, kann ein fairer Vergleich mit dem in Abschnitt 6.2 vorgeschlagenen DNN-basierten RLDS-Schätzer stattfinden.

6.4. Experimentelle Untersuchungen

Als Evaluierungsdaten werden die restlichen $75\,\%$ der Daten des *isolated simulated development* Datensatzes der CHiME-3-Datenbank verwendet, die bei der Optimierung der RLDS-Schätzer nicht gebraucht werden. In diesen Daten, die etwas mehr als zwei Stunden Audiomaterial umfassen, ist jede der vier Störumgebungen *bus*, *caf*, *ped* und *str* in etwa in der gleichen Menge vertreten. Die Evaluierungsdaten weisen die durchschnittlichen eingangsseitigen Werte von $\mathrm{SNR}_{\mathrm{IN}} = 5.6\,\mathrm{dB}$ und MOS-$\mathrm{LQO}_{\mathrm{WB,IN}} = 1.29$ auf. Bei diesen Untersuchungen kommt dasselbe System zur spektralen Entstörung zum Einsatz, welches auch bei der Optimierung der RLDS-Schätzer benutzt wurde. Im Unterschied zu den Optimierungsexperimenten wird hier in allen konventionellen RLDS-Schätzern angenommen, dass in den ersten fünf Rahmen eines gestörten Sprachsignals nur das Störsignal vorhanden ist. Diese Annahme ermöglicht eine bessere Initialisierung der RLDS-Schätzer und erweist sich als vorteilhaft hinsichtlich der Leistungsfähigkeit der Schätzer auf den Evaluierungsdaten. Da die im vorigen Unterkapitel durchgeführte Optimierung der RLDS-Schätzer auch auf den Evaluierungsdaten nicht alle Bewertungsmaße gleichzeitig verbessern konnte, werden



(a) RLDS-Schätzung (b) Sprachsignalentstörung

Abbildung 6.4.: Vergleich des DNN-basierten RLDS-Schätzers mit den zehn modernen Verfahren zur RLDS-Schätzung für die optimierten (opt) und für die empfohlenen (emp) Parameter: (a) Schätzfehler und Schätzfehlervarianz gemessen jeweils in LEM und LEV, (b) Störsignaldämpfung und Qualität der entstörten Sprachsignale gemessen jeweils in SNR_{OUT} und $\text{MOS-LQO}_{\text{WB}}$. Die Evaluierungsdaten weisen die durchschnittlichen eingangsseitigen $\text{SNR}_{\text{IN}} = 5.6\,\text{dB}$ und $\text{MOS-LQO}_{\text{WB,IN}} = 1.29$ auf.

Ergebnisse der konventionellen RLDS-Schätzer sowohl mit den von den Entwicklern empfohlenen (emp) als auch mit den optimierten (opt) Parametern vorgestellt. Für die Berechnung der Messgrößen LEM und LEV wird das Spektrogramm des Störsignals $|D(k,\ell)|^2$ als RLDS-Referenz verwendet. Die resultierenden Bewertungsmaße LEM, LEV, SNR_{OUT} und MOS-LQO gemittelt über die Daten aller vier Störumgebungen sind in Abb. 6.4 dargestellt. Betrachtet man die resultierenden Bewertungsmaße der konventionellen RLDS-Schätzer, fällt auf, dass die durchgeführte Optimierung auch auf den Evaluierungsdaten die Leistungsfähigkeit der modellbasierten RLDS-Schätzer insgesamt steigert.

RLDS-Schätzung: Wie Abb. 6.4 (a) zeigt, führt bessere Initialisierung der RLDS-Schätzer zu den insgesamt kleineren Werten von LEM und LEV im Vergleich zu den Optimierungsergebnissen in Abb. 6.3 (a). Insbesondere profitiert davon das MMSE-VAD-Verfahren, das hier im Unterschied zu den Optimierungsergebnissen unter den besten Verfahren zu finden ist. Während die optimierten MMSE-BM- und SPP-FP-Schätzer unter den konventionellen Verfahren die beste RLDS-Schätzung liefern, schneidet hier am schlechtesten das MCRA-MAP-Verfahren ab. Erfreulicherweise erreicht der DNN-basierte RLDS-Schätzer die beste Genauigkeit unter allen verwendeten RLDS-Schätzern. Im Vergleich zu dem besten modellbasierten MMSE-BM-Schätzer kann der DNN-NPP-Schätzer das LEM-Maß von 5.4 auf 4.6 und das LEV-Maß 19 auf 18.1 reduzieren. Dabei ist zu beachten, dass das verwendete neuronale Netz im Unterschied zu den modellbasierten RLDS-Schätzern keine Daten des *development* Datensatzes sah.

Störsignaldämpfung und Sprachsignalqualität: Gemessen in den Bewertungsmaßen SNR_{OUT} und $\text{MOS-LQO}_{\text{WB}}$ ist die Überlegenheit des DNN-basierten RLDS-Schätzers über die konventionellen Verfahren noch größer, wie Abb. 6.4 (b) zeigt. Hier fällt auf, dass die modellbasierten Verfahren auf Evaluierungsdaten nicht mehr die Leistungsfähigkeit erreichen,

die sie auf den Optimierungsdaten zeigen, vergleiche dazu mit Abb. 6.3 (b). Während unter den modellbasierten Verfahren vom optimierten MMSE-VAD-Schätzer die beste Sprachsignalqualität von MOS-LQO$_{\text{WB,OUT}} = 1.47$ erreicht wird, weist der DNN-basierte Schätzer MOS-LQO$_{\text{WB,OUT}} = 1.51$ auf. Gleichzeitig demonstriert der vorgeschlagene Schätzer eine hervorragende Störsignaldämpfung von SNR$_{\text{OUT}} = 14.2\,\text{dB}$. Die beste Leistungsfähigkeit unter den modellbasierten Verfahren wird hier vom optimierten EMCRA-Verfahren mit SNR$_{\text{OUT}} = 13\,\text{dB}$ erbracht, die jedoch mit den massiven Verlusten in der Sprachsignalqualität verbunden ist. Die Dominanz des DNN-basierten RLDS-Schätzers ist sicherlich auf seine robuste NPP-Schätzung zurückzuführen, die eine stabile genaue RLDS-Schätzung ermöglicht und letztendlich zu einer höheren Leistungsfähigkeit in der spektralen Entstörung führt.

6.5. Zusammenfassung

In diesem Kapitel wurde ein robuster DNN-basierter Maskenschätzer aus [HDCHU15] in einem kausalen Schätzer einer spektralen Rauschleistungsdichte eingesetzt, welcher inspiriert von einigen modellbasierten RLDS-Schätzern auf einer rekursiven Gleichung beruht, die auch in vielen konventionellen RLDS-Schätzern verwendet wird. Dabei wird der vorgeschlagene RLDS-Schätzer vom neuronalen Netz beim Finden der Zeit-Frequenz-Punkte unterstützt, die vom Störsignal dominiert werden. Also liefert das DNN eine kausale Schätzung entsprechender Wahrscheinlichkeiten und zwar allein aus dem Spektrogramm eines gestörten Sprachsignals. Wie aus der Praxis bekannt ist, stellt sich eine solche Aufgabe als sehr herausfordernd dar und ist ein großer Schwachpunkt vieler modellbasierter RLDS-Schätzer. Eingesetzt in einem System zur spektralen Sprachsignalentstörung führte der DNN-basierter RLDS-Schätzer zu einem Hybridsystem, in dem das neuronale Netz mit den konventionellen modellbasierten Verfahren kombiniert wird, die vermutlich zur besseren Generalisierungsfähigkeit des Gesamtsystems beitragen.

Um das Leistungsvermögen des vorgeschlagenen RLDS-Schätzers, dessen DNN auf den Trainingsdaten der CHiME-3-Datenbank trainiert wurde, mit Leistungsfähigkeit konventioneller modellbasierter Verfahren zur RLDS-Schätzung gerecht vergleichen zu können, wurden Parameter der Letzteren auf einem Teil der CHiME-3-Daten optimiert. Als Optimierungsmaß wurde dabei ein kombiniertes Bewertungsmaß verwendet, in welches sowohl die etablierten Bewertungsgrößen einer RLDS-Schätzung als auch die einer Sprachsignalentstörung einfließen. In einem experimentellen Vergleich zeigte sich, dass die Verwendung optimierter Parameter gemittelt über alle betrachteten Verfahren sowohl zu einer genaueren RLDS-Schätzung als auch zu einer qualitativ besseren Signalentstörung führen. Da allerdings einige Bewertungsmaße auf Kosten der anderen verbessert wurden, wurden die konventionellen modellbasierten Verfahren in einer finalen Auswertung sowohl mit dem von den jeweiligen Autoren empfohlenen als auch mit den optimierten Parametern eingesetzt. Dabei wurde gezeigt, dass der DNN-basierte RLDS-Schätzer alle modellbasierten Verfahren in allen betrachteten Bewertungsmaßen übertrifft und zwar unabhängig davon, ob bei den Letzteren ein empfohlener oder ein optimierter Parametersatz verwendet wird.

7. Bayesscher Postprozessor zur RLDS-Schätzung

Viele modellbasierte Verfahren zur RLDS-Schätzung gehen von der legitimen Annahme aus, dass ein Sprachsignal im STFT-Bereich dünn besetzt ist. Somit lassen sich im Kurzzeitspektrum des gestörten Sprachsignals Zeit-Frequenz-Punkte finden, in denen mit guter Näherung nur das Störsignal beobachtet wird und die für die gewünschte RLDS-Schätzung verwendet werden können. Im Gegensatz dazu ist die Unsicherheit der Rauschschätzer während der Sprachaktivität sehr groß. Und da hier die Gefahr, das Sprachsignal als ein Störsignal zu interpretieren, steigt, halten viele Rauschschätzer den LDS-Schätzwert der Störung in solchen Zeit-Frequenz-Punkten auf ihrem letzten sicher geschätzten Wert konstant. Somit wird die zweite Annahme getroffen, nämlich, dass das Störsignal "stationärer" ist als das Sprachsignal und sein LDS sich nicht allzu sehr innerhalb der Sprachaktivität verändert. Diese letzte Annahme wird bei Vorliegen einer nichtstationären Störungen stark verletzt. Somit stellen die nichtstationären Störungen immer noch eine große Herausforderung für moderne RLDS-Schätzer dar, denn ihre zeitvariante spektrale Leistungsdichte bleibt auch während der Sprachaktivität nicht konstant und bedarf einer Verfolgung [RL06]. Dies ist besonders wichtig für die Zeit-Frequenz-Punkte, in denen die Leistungsdichte der Störung vom ungestörten Sprachsignal nicht ausreichend maskiert oder verdeckt wird. Der hier vorgestellte Bayessche Schätzer soll die RLDS-Schätzung während Sprachaktivität verbessern, indem er als Postprozessor zur ersten Stufe eines Systems zur Signalentstörung arbeitet und die RLDS-Schätzung der ersten Stufe verfeinert. Ein innovativer Ansatz dafür liegt im Paradigmenwechsel, die Welt mit den Augen des Rauschschätzers zu sehen, für den das Sprachsignal störend ist und ihm seine Arbeit besonders während der Sprachaktivität erschwert.

In einem System zur spektralen Sprachsignalentstörung wie in Abb. 2.2 wird zunächst das RLDS geschätzt, aus dem anschließend das *a priori* SNR, die *a posteriori* SPP und schließlich die spektrale Filterfunktion berechnet werden. Der Rauschschätzer dient dabei als ein Sprungbrett zum Erfolg weiterer Systemkomponenten. Nach dem Prinzip der Nächstenliebe könnten aber auch diese Komponenten die RLDS-Schätzung in ihrer Arbeit unterstützen und zwar in besonders schwierigen Regionen mit oben erwähnter Sprachaktivität. Dafür kann ein System mit zwei hintereinander geschalteten Stufen verwendet werden. In diesem würde der RLDS-Schätzer der ersten Stufe eine grobe RLDS-Schätzung liefern und der Rauschschätzer der zweiten Stufe, der als ein Postprozessor agiert, die verfeinerte RLDS-Schätzung berechnen, die auch während der Sprachaktivität die RLDS schätzt. Während der erste RLDS-Schätzer nur aus dem gestörten Eingangssignal seine Schätzwerte berechnet, soll der RLDS-Postprozessor zusätzlich von den anderen Systemkomponenten der ersten Stufe unterstützt werden. Die Schätzwerte des zu entwickelnden Postprozessors werden anschließend in der zweiten Stufe für die Signalentstörung eingesetzt.

7.1. MAP-basierter Postprozessor

Die statistische Formulierung des Schätzproblems, das vom zu entwickelnden Postprozessor gelöst werden soll, lautet: Schätze die Varianz $\lambda_D(k,\ell)$ eines komplexwertigen mittelwertfreien nichtstationären normalverteilten Rauschens $D(k,\ell)$ aus den Beobachtungen des gestörten Zufallsprozesses $Y(k,\ell)$, wobei aus Sicht des Postprozessors die STFT-Koeffizienten des ungestörten Sprachsignals $S(k,\ell)$ als eine additive komplexwertige mittelwertfreie nichtstationäre normalverteilte Störung mit einer bekannten zeitvarianten Varianz $\lambda_S(k,\ell)$ aufzufassen sind, die von der ersten Systemstufe geschätzt wird[1]. Dabei soll die Varianz $\lambda_D(k,\ell)$ nicht als ein unbekannter Parameter modelliert werden, sondern als eine Zufallsgröße. Diese statistische Modellierung ist in den Gleichungen von (2.3) bis (2.9) verankert. Während eine MAP-basierte Lösung dieses Problems für einen reellwertigen Zufallsprozess in [KHU11] beschrieben wird, widmet sich dieses Unterkapitel einer MAP-basierten Lösung für den Fall von komplexwertigen Zufallsprozessen. Die Herleitung dieser Lösung wird in Unterabschnitt 7.1.1 vorgestellt und führt zum gewünschten MAP-basierten RLDS-Schätzers, der in Unterabschnitt 7.1.2 als ein Postprozessor in einem zweistufigen System zur einkanaligen akustischen Signalentstörung eingesetzt wird. Die experimentellen Ergebnisse der durchgeführten Machbarkeitsstudie zeigen dabei, dass der hergeleitete Postprozessor die RLDS-Schätzung der ersten Systemstufe wirklich verbessern kann. Der in diesem Abschnitt vorgestellte MAP-basierte (MAPB) Postprozessor wurde zum ersten Mal in [CKTVHU12] in seiner nichtoptimierten Version vorgestellt.

7.1.1. Herleiten des Bayesschen Schätzers

Der zu entwickelnde MAP-basierte RLDS-Schätzer wird sukzessive in drei Schritten hergeleitet. Im ersten Schritt wird ein Spezialfall betrachtet, bei dem der Zufallsprozess $D(k,\ell)$ stationär und das Sprachsignal abwesend $Y(k,\ell) = D(k,\ell)$ sind. Im zweiten Schritt wird die Annahme einer stationären Störung fallen gelassen, sodass man nichtstationäre Störungen zulässt. Im dritten Schritt betrachtet man den allgemeinen Fall, dass eine nichtstationäre Störung vom Sprachsignal additiv überlagert wird $Y(k,\ell) = D(k,\ell) + S(k,\ell)$. Da der MAP-basierte RLDS-Schätzer in jedem Frequenzband dieselbe Signalverarbeitung durchführen soll, wird in weiteren Herleitungen der Frequenzindex k der Übersichtlichkeit halber weggelassen. Um die Formeln kompakt zu halten, wird der Zeitindex ℓ tief gesetzt.

Erster Schritt: Zunächst wird der Spezialfall betrachtet, bei dem keine Sprachaktivität vorliegt, $Y_\ell = D_\ell$, und der Zufallsprozess D_ℓ sich als eine stationäre Störung mit einer zeitunabhängigen Varianz $\lambda_{D,\ell} = \lambda_{D,\ell-1} = \lambda_D$ darstellt. Ähnlich wie in Abschnitt 5.2 wird auch hier das unbekannte RLDS λ_D als Zufallsvariable modelliert, sodass die Verteilungsdichtefunktion (VDF) einer komplexwertigen Beobachtung Y_ℓ zu einer bedingten VDF mit der Form wird:

$$p_{Y_\ell|\lambda_D,H_0}(y_\ell|\lambda) = \frac{1}{\pi\cdot\lambda}\cdot\exp\left(-\frac{|y_\ell|^2}{\lambda}\right), \tag{7.1}$$

wobei λ eine Realisierung von λ_D ist. Da λ_D hier im Unterschied zur Gleichung (2.31) eine versteckte Zufallsvariable ist, wird sie von der beobachtbaren Zufallsvariable Y_ℓ mit

[1] Da $S(k,\ell)$ und $D(k,\ell)$ als mittelwertfrei angenommen werden, spielen die Leistungsdichtespektren $\lambda_S(k,\ell)$ und $\lambda_D(k,\ell)$ die Rolle der Varianzen der jeweiligen Verteilungen in (2.4) und (2.5).

einem senkrechten Strich getrennt. Genauso wie in Abschnitt 5.2 wird λ_D auch hier mit einer skalierten inversen Chi-Quadrat-Verteilung modelliert

$$p_{\lambda_D}(\lambda) = \text{Inv-}\chi^2(\lambda_D; \nu_\ell, \tau_\ell^2), \tag{7.2}$$

die in (5.9) bereits definiert wurde. Dabei sind $\nu_\ell > 0$ ein Freiheitsgrad und $\tau_\ell^2 > 0$ ein Skalierungsparameter. Die VDF (7.2) ist für den betrachteten Spezialfall eine konjugierte *a priori* Verteilung der Beobachtungsverteilung (7.1) bezüglich der versteckten Zufallsvariablen λ_D und bringt mit sich das Vorwissen über die Zufallsvariable λ_D, das vor der Verarbeitung der Beobachtung y_ℓ im LDS-Rauschschätzer vorliegt. Dieses Vorwissen wird aus der Verarbeitung von Daten der vorigen Rahmen $\{y_1, ..., y_{\ell-1}\}$ gewonnen und ist in den Hyperparametern ν_ℓ und τ_ℓ^2 gespeichert. Das *a priori* Wissen des MAP-basierten RLDS-Schätzers hängt also von den Daten ab, die er bereits gesehen hat.[2] Im LDS-Bereich ist dies auch zu erwarten, denn die Beobachtungen der aufeinander folgenden Rahmen sind nicht unkorreliert (auch nicht für weißes Rauschen) und zwar aufgrund der Signalverarbeitung der STFT mit überlappenden Fenstern.

Mit der Bayesschen Regel für bedingte Wahrscheinlichkeiten kann aus den Gleichungen (7.1) und (7.2) die *a posteriori* Verteilungsdichtefunktion ausgerechnet werden

$$p_{\lambda_D|Y_\ell}(\lambda|y_\ell; \tilde{\nu}_\ell, \tilde{\tau}_\ell^2) = \text{Inv-}\chi^2(\lambda_D; \tilde{\nu}_\ell, \tilde{\tau}_\ell^2), \tag{7.3}$$

die eine inverse skalierte Chi-Quadrat-Verteilung mit den Hyperparametern $\tilde{\nu}_\ell$ and $\tilde{\tau}_\ell^2$ ist, die mit Hilfe der Aktualisierungsgleichungen wie (5.10) und (5.11) berechnet werden

$$\tilde{\nu}_\ell = \nu_\ell + 2, \tag{7.4}$$

$$\tilde{\tau}_\ell^2 = \frac{\nu_\ell \cdot \tau_\ell^2 + 2 \cdot |y_\ell|^2}{\nu_\ell + 2}. \tag{7.5}$$

Für den MAP-basierten RLDS-Schätzwert $\hat{\lambda}_{D,\ell}$ im ℓ-ten Rahmen resultiert dann

$$\hat{\lambda}_{D,\ell} = \underset{\lambda}{\text{argmax}}\ p_{\lambda_D|Y_\ell}(\lambda|y_\ell; \tilde{\nu}_\ell, \tilde{\tau}_\ell^2) \ = \frac{\tilde{\nu}_\ell}{\tilde{\nu}_\ell + 2} \cdot \tilde{\tau}_\ell^2. \tag{7.6}$$

Setzt man die Aktualisierungsgleichungen (7.4) und (7.5) in (7.6) ein und berücksichtigt, dass der RLDS-Schätzwert $\hat{\lambda}_{D,\ell-1}$ im Rahmen $\ell - 1$ auch ein MAP-Schätzwert ist, der über die Hyperparameter der aktuellen *a priori* Verteilung zu $\hat{\lambda}_{D,\ell-1} = \frac{\nu_\ell}{\nu_\ell+2} \cdot \tau_\ell^2$ ausgerechnet wird, kann der aktuelle Schätzwert wie folgt geschrieben werden

$$\hat{\lambda}_{D,\ell} = \frac{\tilde{\nu}_\ell}{\tilde{\nu}_\ell + 2} \cdot \hat{\lambda}_{D,\ell-1} + \frac{2}{\tilde{\nu}_\ell + 2} \cdot |y_\ell|^2. \tag{7.7}$$

Die Art und Weise, wie $\hat{\lambda}_{D,\ell-1}$ berechnet wird, realisiert den oben erwähnten Aspekt, dass das *a priori* Wissen für jeden Datenblock ℓ aktualisiert wird. Dies geschieht dadurch, dass die *a posteriori* Verteilung, die auf der Beobachtung des vorigen Datenblocks berechnet wird, als *a priori* Verteilung für den aktuellen Datenblock verwendet wird

$$p_{\lambda_D}(\lambda; \nu_\ell, \tau_\ell^2) = p_{\lambda_D|Y_{\ell-1}}(\lambda|y_{\ell-1}; \tilde{\nu}_{\ell-1}, \tilde{\tau}_{\ell-1}^2). \tag{7.8}$$

[2] An dieser Stelle soll erwähnt werden, dass es Schätzverfahren gibt, bei denen das *a priori* Wissen nicht von Beobachtungen abhängig ist. Im MAP-basierten LDS-Schätzer ist dies eindeutig nicht der Fall.

Die Hyperparameter werden demnach folgendermaßen zugewiesen $\nu_\ell = \tilde{\nu}_{\ell-1}$ und $\tau_\ell^2 = \tilde{\tau}_{\ell-1}^2$.

An dieser Stelle lohnt es sich, die Gleichung (7.7) etwas genauer anzuschauen. So sieht man darin, dass der aktuelle MAP-Schätzwert $\hat{\lambda}_{D,\ell}$ sich als eine gewichtete Summe des alten MAP-Schätzwertes $\hat{\lambda}_{D,\ell-1}$ und des Betragsquadrates der aktuellen Beobachtung $|y_\ell|^2$ darstellt. Die entsprechenden Gewichte summieren sich dabei zu 1 auf. Somit bestimmt der Freiheitsgrad $\tilde{\nu}_\ell$ das Gewicht, mit dem das Betragsquadrat der Beobachtung $|y_\ell|^2$ in den MAP-Schätzwert $\hat{\lambda}_{D,\ell}$ einfließt. Gleichzeitig hängt auch die Varianz des MAP-Schätzers in (7.7) vom Freiheitsgrad $\tilde{\nu}_\ell$ ab. Je größer der Wert von $\tilde{\nu}_\ell$, desto kleiner wird die Varianz des Schätzers, denn die Gleichung (7.7) kann auch als eine rekursive Glättung der aufeinander folgenden Beobachtungen $\{\ldots\ |y_{\ell-1}|^2,\ |y_\ell|^2,\ |y_{\ell+1}|^2\ \ldots\}$ angesehen werden

$$\hat{\lambda}_{D,\ell} = \alpha_\ell \cdot \hat{\lambda}_{D,\ell-1} + (1-\alpha_\ell) \cdot |y_\ell|^2. \tag{7.9}$$

Dabei ist $\alpha_\ell = \frac{\tilde{\nu}_\ell}{\tilde{\nu}_\ell+2} \in (0;1)$ ein zeitvarianter Glättungsparameter, der die Bandbreite dieses digitalen IIR-Tiefpassfilters bestimmt.[3] Wird der Hyperparameter $\tilde{\nu}_\ell$ entsprechend der Aktualisierungsgleichung (7.4) fortlaufend berechnet, resultiert daraus $\tilde{\nu}_\ell = \tilde{\nu}_{\ell-1} + 2$. Somit wächst der Wert von $\tilde{\nu}_\ell$ ständig an und führt dazu, dass der Glättungsparameter gegen 1 strebt, d. h. die Bandbreite des digitalen Tiefpassfilters wird schmaler und die Varianz des MAP-Schätzers sinkt. In Abwesenheit des Sprachsignals und im Fall einer stationären Störung ist ein solches Verhalten des LDS-Rauschschätzers logisch und sogar erwünscht, jedoch nicht, wenn eine nichtstationäre Störung vorliegt.

Zweiter Schritt: Im zweiten Schritt der Herleitung des MAP-basierten RLDS-Schätzers wird die Annahme fallen gelassen, dass das Störsignal D_ℓ stationär ist. Es wird also eine nichtstationäre Störung mit einer zeitveränderlichen Varianz $\lambda_{D,\ell} \neq \lambda_{D,\ell-1}$ vorausgesetzt. Die Annahme der Abwesenheit des Sprachsignals, $S_\ell = 0$, bleibt dagegen vorerst noch erhalten. Dieser Fall entspricht der Schätzung einer nichtstationären Störung in den Zeit-Frequenz-Punkten ohne Sprachaktivität. Hier ist die rekursive Glättung (7.9) wirklich sehr beliebt in der RLDS-Schätzung und wird bei vielen konventionellen RLDS-Schätzern eingesetzt, siehe Abschnitt 3.1. Die praktische Bewandtnis der Glättung (7.9), die aus der Modellierung der Varianz als eine Zufallsvariable hervorgeht, ist indirekt eine Legitimation für die Wahl der skalierten inversen Chi-Quadrat-Verteilung als konjugierte *a priori* Verteilung.

Eine Reihe konventioneller RLDS-Schätzer verwendet einen konstanten Glättungsparameter $\alpha_{\ell+1} = \alpha_\ell = \alpha_0$ [Hir93, Mar94, Dob95, Yu09]. So wird auch im MAP-basierten RLDS-Schätzer zunächst ein konstanter Glättungsparameter benutzt, der aus einem konstanten Freiheitsgrad hervorgeht

$$\nu_{\ell+1} = \nu_\ell = \nu_0. \tag{7.10}$$

Die Gleichung (7.10) ist nichts anderes als eine Modifikation der Aktualisierungsgleichung (7.4). Diese Modifikation ermöglicht die RLDS-Schätzung einer nichtstationären Störung. Mit dem konstanten Freiheitsgrad ν_0 verändert sich das Gewicht der aktuellen Beobachtung nicht mehr. Auf diese Weise wird dem RLDS-Schätzer die wichtige Eigenschaft hinzugefügt, die weiter in der Vergangenheit liegende Beobachtungen zu vergessen. Die Wahl von ν_0 hängt davon ab, wie der Kompromiss zwischen der Varianz des RLDS-Schätzers in Gegenwart einer stationären Störung und seiner Verfolgungsfähigkeit der zeitlich veränderlichen Varianz $\lambda_{D,\ell}$ gelöst wird. Man beachte, dass im Verlauf dieses Kapitels eine weitere

[3] Der Frequenzgang dieses einfachen digitalen IIR-Tiefpassfilters kann in Abhängigkeit vom Freiheitsgrad $\tilde{\nu}_\ell$ leicht berechnet werden.

Möglichkeit vorgestellt wird, wie man den Freiheitsgrad ν_ℓ in der Gegenwart einer nichtstationären Störung anders als mit der Gleichung (7.4) steuert und somit doch zeitvariant lässt.

Dritter Schritt: Im dritten Schritt wird die Annahme der Sprachsignalabwesenheit fallen gelassen, $\lambda_{S,\ell} \neq 0$, wodurch zum ersten Mal ein praxisrelevanter Fall entsteht, bei dem die Beobachtungsverteilung (7.1) folgendermaßen umgeschrieben wird

$$p_{Y_\ell|\lambda_D}(y_\ell|\lambda;\lambda_{S,\ell}) = \frac{1}{\pi\cdot(\lambda_{S,\ell}+\lambda)}\cdot\exp\left(-\frac{|y_\ell|^2}{\lambda_{S,\ell}+\lambda}\right). \tag{7.11}$$

Dabei wird vorausgesetzt, dass das zeitinvariante LDS des ungestörten Sprachsignals $\lambda_{S,\ell}$ bekannt ist. Aus der Bayesschen Regel für bedingte Wahrscheinlichkeiten resultiert dann folgender funktionaler Zusammenhang für die *a posteriori* Verteilungsdichtefunktion

$$p_{\lambda_D|Y_\ell}(\lambda|y_\ell;\nu_0,\tau_\ell^2,\lambda_{S,\ell}) \;\propto\; \frac{(\lambda)^{-\frac{\nu_0+2}{2}}}{\lambda_{S,\ell}+\lambda}\cdot\exp\left(-\frac{|y_\ell|^2}{\lambda_{S,\ell}+\lambda}-\frac{\nu_0\tau_\ell^2}{2\lambda}\right), \tag{7.12}$$

die keine skalierte inverse Chi-Quadrat-Verteilung mehr ist. Dadurch verliert (7.2) ihre Eigenschaft einer konjugierten *a priori* Verteilung. Diese Tatsache verkompliziert die angestrebte MAP-Schätzung sehr und macht sie rechnerisch sehr ineffizient bei einer Realisierung in einem Datenstrom. Eine Wiederherstellung der wichtigen Eigenschaft einer konjugierten *a priori* Verteilung wäre von daher hinsichtlich der Implementierung des LDS-Schätzers von großem Vorteil. Zunächst allerdings soll die Maximumstelle $\hat{\lambda}_{D,\ell}$ der *a posteriori* VDF in (7.12) gefunden werden, die ja zum MAP-basierten RLDS-Schätzer (7.6) führt.

Das Maximum von (7.12) entspricht der Minimumstelle der Funktion:

$$\begin{aligned} f(\lambda) &= -2\cdot\ln(\,p_{\lambda_D|Y_\ell}(\,\lambda\,|y_\ell;\nu_0,\tau_\ell^2,\lambda_{S,\ell})\,) \\ &\propto \underbrace{2\cdot\ln(\lambda_{S,\ell}+\lambda)+(\nu_0+2)\cdot\ln(\lambda)}_{f_1(\lambda)} + \underbrace{\frac{2|y_\ell|^2}{\lambda_{S,\ell}+\lambda}+\frac{\nu_0\tau_\ell^2}{\lambda}}_{f_2(\lambda)}. \end{aligned} \tag{7.13}$$

Dabei werden Funktionen $f_1(\lambda)$ und $f_2(\lambda)$ so definiert, dass $f_1(\lambda)$ monoton steigend und $f_2(\lambda)$ monoton fallend für $\lambda > 0$ ist. Und da es $\lim_{\lambda\to 0} f(\lambda) = \lim_{\lambda\to\infty} f(\lambda) = \infty$ gilt, hat $f(\lambda)$ eine einzige Minimumstelle, die als positive Nullstelle der ersten Ableitung der Funktion $f(\lambda)$ gefunden werden kann. Diese Ableitung lässt sich wie folgt berechnen:

$$f'_\lambda(\lambda) = \frac{2}{\lambda_{S,\ell}+\lambda}+\frac{\nu_0+2}{\lambda}-\frac{2|y_\ell|^2}{(\lambda_{S,\ell}+\lambda)^2}-\frac{\nu_0\tau_\ell^2}{\lambda^2} \tag{7.14}$$

$$= \frac{(\nu_0+2)(\lambda-\lambda_a)(\lambda_{S,\ell}+\lambda)^2+2(\lambda-\lambda_b)\lambda^2}{(\lambda_{S,\ell}+\lambda)^2\cdot\lambda^2} \tag{7.15}$$

mit zwei markanten Punkten

$$\lambda_a = \frac{\nu_0}{\nu_0+2}\cdot\tau_\ell^2 > 0 \quad\text{und}\quad \lambda_b = |y_\ell|^2-\lambda_{S,\ell}. \tag{7.16}$$

Da der Nenner von (7.15) immer positiv ist, reicht es für die Nullstellensuche, nur den Zähler zu betrachten, der im Weiteren als Funktion

$$g(\lambda) = (\nu_0+2)(\lambda-\lambda_a)(\lambda_{S,\ell}+\lambda)^2+2(\lambda-\lambda_b)\lambda^2 \tag{7.17}$$

bezeichnet wird. Unter der Annahme $\lambda_{S,\ell} < |y_\ell|^2$ und somit $\lambda_b > 0$, gelten für das Polynom des dritten Grades $g(\lambda)$ folgende Ungleichungen:

$$g(\lambda_{\mathrm{U}}) < 0 \quad \text{für die Untergrenze} \quad \lambda_{\mathrm{U}} = \min(\lambda_a, \lambda_b) > 0 \tag{7.18}$$

$$g(\lambda_{\mathrm{O}}) > 0 \quad \text{für die Obergrenze} \quad \lambda_{\mathrm{O}} = \max(\lambda_a, \lambda_b) > 0. \tag{7.19}$$

Die gesuchte Nullstelle $\hat{\lambda}_{D,\ell}$ liegt somit im Intervall $[\lambda_{\mathrm{U}}; \lambda_{\mathrm{O}}]$ und kann effizient mit Kombination des Intervallhalbierungsverfahrens und des Newton Verfahrens nummerisch berechnet werden. Bei einer genaueren Analyse des MAP-basierten RLDS-Schätzwertes $\hat{\lambda}_{D,\ell}$ fällt auf, dass $\hat{\lambda}_{D,\ell}$ sich stets zwischen den markanten Punkten λ_a und λ_b befindet. Während der Punkt λ_a allein auf dem *a priori* Vorwissen basiert, fließt in den Punkt λ_b die aktuelle Beobachtung ein. Außerdem könnte man den Punkt λ_b auch als einen RLDS-Schätzer betrachten, der mittels spektraler Subtraktion berechnet wird. Der resultierende Schätzwert $\hat{\lambda}_{D,\ell}$ ist also eine Kombination aus dem *a priori* Wissen und aus der aktuellen Beobachtung. Im Fall $\lambda_{S,\ell} = 0$ führt die Lösung der Gleichung $g(\lambda) = 0$ wie erwartet auf die rekursive Gleichung (7.7) mit dem konstanten Freiheitsgrad $\nu_\ell = \nu_0$. Der Sonderfall $\lambda_a = \lambda_b$ geht mit einer stationären Störung $\hat{\lambda}_{D,\ell} = \hat{\lambda}_{D,\ell-1}$ einher.

Nachdem der gesuchte MAP-Schätzwert $\hat{\lambda}_{D,\ell}$ gefunden wurde, bleibt nur noch die Frage zu beantworten, wie das Vorwissen des RLDS-Schätzers für den nächsten Rahmen aktualisiert werden soll, das die aus der Beobachtung y_ℓ gewonnenen Informationen berücksichtigt. Die Verwendung der wahren *a posteriori* VDF $p_{\lambda_D|Y_\ell}(\lambda|y_\ell; \nu_\ell, \tau_\ell^2, \lambda_{S,\ell})$ aus (7.12) würde dafür sorgen, dass die *a priori* Verteilungen des LDS-Schätzers für unterschiedliche Rahmen nicht aus ein und derselben Verteilungsfamilie stammen. Selbst für die VDF in (7.12) ist es schwierig zu bestimmen, aus welcher Verteilungsfamilie sie stammt. Diese Inkonsistenz würde die Recheneffizienz des MAP-basierten RLDS-Schätzers ruinieren.

Um die Recheneffizienz des RLDS-Schätzers aufrecht zu erhalten, wird die Approximation der wahren *a posteriori* Verteilung in (7.12) durch eine inverse skalierte Chi-Quadrat-

Algorithm 7.1 MAP-basierter RLDS-Schätzer mit einem konstanten Freiheitsgrad

Input: Spektrogramm $|y_{k,\ell}|^2$ und Schätzwerte $\breve{\lambda}_{S,k,\ell}$ für $k \in [1;\, K]$ und $\ell \in [1;\, L]$
Parameter: konstanter Freiheitsgrad ν_0
Output: RLDS-Schätzwert des Störsignals $\hat{\lambda}_{D,k,\ell}$

```
loop
  Für jeden Frequenzband k ∈ [1; K] berechne
  for ℓ = 1 to L do
    if ℓ == 1 then
      den initialen RLDS-Schätzwert λ̂_{D,k,1} = (1/2) · |y_{k,1}|^2
    else
      die markanten Punkte λ_a = (ν_0/(ν_0+2)) · λ̂_{D,k,ℓ−1} und λ_b = |y_{k,ℓ}|^2 − λ̆_{S,k,ℓ} > 0
      die Nullstelle λ̂_{D,k,ℓ} von g_{k,ℓ}(λ) = (ν_0 + 2)(λ − λ_a)(λ̆_{S,k,ℓ} + λ)^2 + 2(λ − λ_b)λ^2 in
        Intervallgrenzen λ̂_{D,k,ℓ} ∈ (λ_U; λ_O) mit λ_U = min(λ_a, λ_b) und λ_O = max(λ_a, λ_b)
    end if
    den Skalierungsparameter für den nächsten Rahmen τ^2_{k,ℓ} = ((ν_0+2)/ν_0) · λ̂_{D,k,ℓ}
  end for
end loop
```

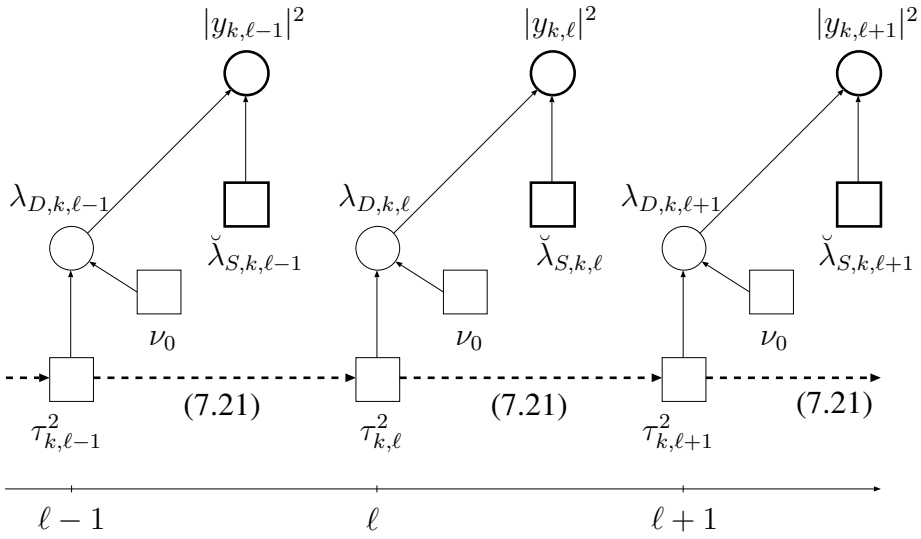


Abbildung 7.1.: Graphische Darstellung der statistischen Zusammenhänge des MAP-basierten Postprozessors im k-ten Frequenzband.

Verteilung vorgeschlagen:

$$p_{\lambda_D}(\lambda;\nu_0,\tau^2_{\ell+1}) = \text{Inv-}\chi^2(\lambda;\tilde{\nu}_0,\tilde{\tau}^2_\ell) \approx p_{\lambda_D\,|\,Y_\ell}(\lambda|y_\ell;\nu_0,\tau^2_\ell,\lambda_{S,\ell}). \tag{7.20}$$

Dabei soll die approximierte VDF denselben Modus der wahren *a posteriori* Verteilung bekommen, der unter Berücksichtigung der Gleichung (7.10) durch folgende Wahl des Skalierungsparameters realisiert wird

$$\tau^2_{\ell+1} = \tilde{\tau}^2_\ell = \frac{\nu_0+2}{\nu_0}\cdot\hat{\lambda}_{D,\ell}. \tag{7.21}$$

Somit wird dafür gesorgt, dass die *a priori* Verteilung des MAP-basierten RLDS-Schätzers für jeden Datenblock eine inverse skalierte Chi-Quadrat-Verteilung ist, und als Folge, behält der Schätzer seine Recheneffizienz auf dem Niveau eines konventionelles MAP-Schätzer mit einer konjugierten *a priori* Verteilung. Man beachte, dass der MAP-basierte Postprozessor immer noch den wahren MAP-Schätzwert berechnet, sodass die eingeführte Approximation nur für die Aktualisierung des *a priori* Wissens verwendet wird. Neben der Recheneffizienz hat der MAP-basierte LDS-Schätzer einen weiteren erwähnenswerten Vorteil, der darin besteht, dass er einen einzigen zeitinvarianten Parameter besitzt und zwar den Freiheitsgrad ν_0, welcher die Verfolgungseigenschaften (oder die Bandbreite) des Schätzers bestimmt und entsprechend der Nichtstationarität des Störsignals gewählt werden sollte.

Die Berechnungsschritte des MAP-basierten RLDS-Schätzers werden in Alg. 7.1 zusammengefasst. Dabei wird statt des wahren Wertes des LDS des ungestörten Sprachsignals $\lambda_{S,k,\ell}$, der in der Praxis unbekannt ist, sein Schätzwert $\breve{\lambda}_{S,k,\ell}$ verwendet. Wie dieser Schätzwert berechnet wird, wird im Weiteren ausführlich erläutert. Die graphische Darstellung der statistischen Zusammenhänge des MAP-basierten Postprozessors im k-ten Frequenzband ausgerollt über die Zeitachse ist in Abb. 7.1 präsentiert. Dabei sind die Zufallsvariablen durch Kreise und die Parameter durch Rechtecke gekennzeichnet. Beobachtbare Zufallsvariablen und gegebene Parameter sind durch Kreise und Rechtecke mit dicken Linien gekennzeichnet, entsprechend der Notation in [Bis06]. Durch die Verbindungen der Skalierungspa-

rameter der aufeinander folgenden Rahmen, die mit Hilfe von (7.21) realisiert werden, wird die Aktualisierung des *a priori* Wissens verdeutlicht.

7.1.2. MAP-basierter RLDS-Schätzer als Postprozessor

Bei der Herleitung der Gleichungen des MAP-basierten Postprozessors wurde angenommen, dass die spektrale Leistungsdichte des ungestörten Sprachsignals $\lambda_{S,k,\ell}$ in jedem Zeit-Frequenz-Punkt gegeben ist. In Wirklichkeit ist sie allerdings unbekannt und muss geschätzt werden. Aus diesem Grund benötigt der hergeleitete MAP-basierte Schätzer eine Vorverarbeitungsstufe, zu der er als ein Postprozessor agiert[4]. In dieser ersten Stufe kann das konventionelle DD-Verfahren eingesetzt werden [EM84], das den Schätzwert des *a priori* SNR $\breve{\xi}_{k,\ell}$ bereitstellt, aus dem anschließend der gewünschte Schätzwert $\breve{\lambda}_{S,k,\ell}$ für den MAP-basierten Postprozessor berechnet wird. Da das DD-Verfahren seinerseits einen Schätzwert der Rauschleistungsdichte $\breve{\lambda}_{D,k,\ell}$ benötigt, muss ein konventioneller RLDS-Schätzer dem DD-Verfahren vorgeschaltet werden. In den experimentellen Untersuchungen dieses Unterkapitels wird das IMCRA-Verfahren als ein RLDS-Schätzer verwendet [Coh03]. Fügt man zum IMCRA- und DD-Verfahren noch ein OMLSA-Verfahren zur SPP-basierten Berechnung der spektralen Filterfunktion $\breve{G}_{k,\ell}$ hinzu [CB01b], erhält man ein konventionelles System zur Entstörung der Sprachsignale, siehe Abb. 7.2. Das in der ersten Stufe geschätzte Betragsspektrum des ungestörten Signals wird im Weiteren mit $|\breve{S}_{k,\ell}|$ bezeichnet.

Diese erste Systemstufe soll als ein Vergleichssystem für die ersten experimentellen Untersuchungen des MAP-basierten Postprozessors dienen, der als ein RLDS-Schätzer im nachgeschalteten System zur Signalentstörung eingesetzt wird. Aus den Schätzwerten des MAPB-Postprozessors $\hat{\lambda}_{D,k,\ell}$ können anschließend mit DD- und OMLSA-Verfahren zunächst das *a priori* SNR $\hat{\xi}_{k,\ell}$ und danach die spektrale Filterfunktion $\hat{G}_{k,\ell}$ berechnet werden, die für die

[4]In [TLA09] wird ein System zur Entstörung der Sprachsignale mit einer zweistufigen Rauschschätzung vorgestellt. Die Rauschschätzung mit Verwendung des MAP-basierten Postprozessors ist ja auch zweistufig.

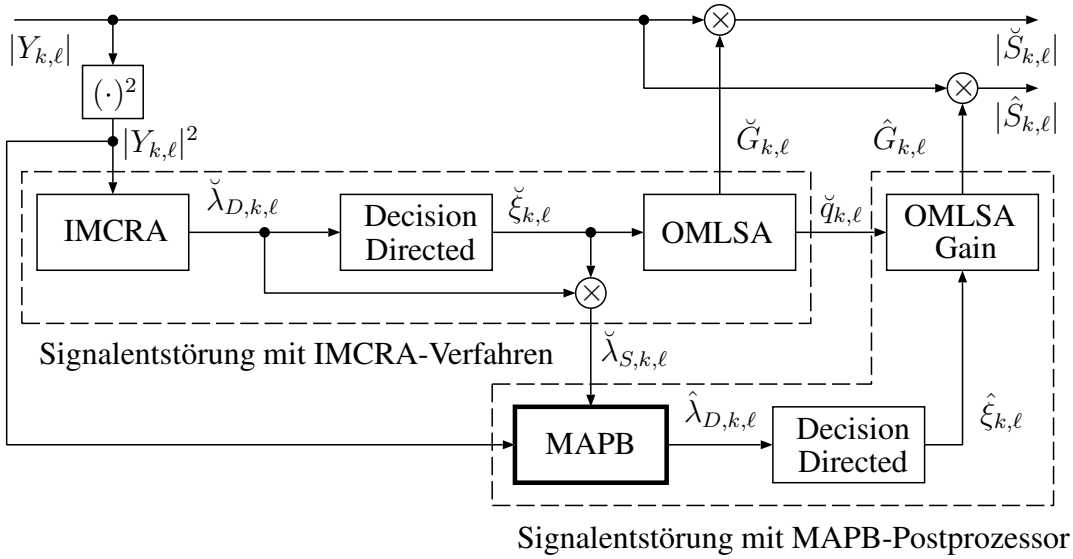


Abbildung 7.2.: Zweistufiges System zur Entstörung spektraler Amplituden unter Verwendung des IMCRA-Algorithmus und des MAPB-Postprozessors zur RLDS-Schätzung.

Berechnung der spektralen Amplitude in der zweiten Systemstufe $|\hat{S}_{k,\ell}|$ verwendet wird. Die Untersuchungen haben gezeigt, dass Schätzwerte $\breve{q}_{k,\ell}$ der Wahrscheinlichkeit für Sprachsignalabwesenheit beim OMLSA-Verfahrens der ersten Systemstufe für die Signalentstörung bereits gut genug sind und in der zweiten Stufe nicht noch einmal berechnet werden müssen. Somit darf die spektrale Filterfunktion des OMLSA-Verfahrens in der zweiten Stufe die Schätzwerte $\breve{q}_{k,\ell}$ verwenden. Man beachte, dass die OMLSA-Filterfunktion die RLDS-Schätzwerte der jeweiligen Stufe verwendet, die der Übersichtlichkeit halber in Abb. 7.2 nicht explizit gekennzeichnet sind. Sonst wird für die Berechnung der entstörten Zeitsignale der beiden Systemstufen $\breve{s}(t)$ und $\hat{s}(t)$ die Phase des gestörten Spektrums $Y_{k,\ell}$ verwendet, wie in Abb. 2.1 angedeutet.

Im Rahmen einer Machbarkeitsstudie für den hergeleiteten MAP-basierten Postprozessor werden erste experimentelle Untersuchungen mit ungestörten Aufnahmen aus der TIMIT-Datenbank durchgeführt, siehe Abschnitt 2.4. Bevor die einzelnen ungestörten Audioaufnahmen zu den längeren 3-minütigen Signalen verbunden werden, werden Pausen am Anfang und am Ende jeder Äußerung entfernt. Auf diese Weise werden separate Aufnahmen mit weiblichen und mit männlichen Sprechern erzeugt, in denen Sprachsignale von jeweils 7 unterschiedlichen Personen vorkommen. Anschließend werden die ungestörten Aufnahmen künstlich mit 4 verschiedenen Arten von Störsignalen additiv überlagert. Dabei werden folgende Störsignale der SPIB-Datenbank verwendet: das stationäre weiße normalverteilte Rauschen (WNR), der in der SPIB-Datenbank als Rauschtyp *white* bezeichnet wird, und die Störsignale der Rauschtypen *babble* und *factory-1*. Um die Leistungsfähigkeit des MAPB-Postprozessors beim stark nichtstationären Rauschen zu untersuchen, wird zusätzlich ein amplitudenmoduliertes weißes Rauschen mit abwechselnd steigendem und fallendem Rauschpegel generiert, das im Weiteren als nichtstationäres weißes normalverteiltes Rauschen (NWNR) bezeichnet wird. Dabei wird der Rauschpegel im Signal mit der Rate von $2\,\text{dB/s}$ innerhalb von 3 Sekunden zunächst erhöht und dann mit derselben Rate auf das Ausgangsniveau abgesenkt. Die Experimente werden für gestörte Signale bei unterschiedlichen SNR_{IN}-Werten durchgeführt. Dabei wird das globale SNR_{IN} von $-5\,\text{dB}$ bis auf $15\,\text{dB}$ in $5\,\text{dB}$ Schritten erhöht. Alle Signale sind mit der Frequenz von $16\,\text{kHz}$ abgetastet. Bei der STFT wird das Hamming-Analysefenster und die FFT-Länge von 512 mit dem Rahmenvorschub von $25\,\%$ verwendet. Der konstante Freiheitsgrad des MAPB-Postprozessors wird in diesen Experimenten zu $\nu_0 = 40$ gesetzt. Die restlichen Systemkomponenten werden mit den Parameterwerten eingesetzt, die in [Coh03] empfohlen werden. Als RLDS-Referenz $\lambda_{D,k,\ell}$, die mit $\lambda_{D,k,0} = |D_{k,0}|^2$ initialisiert wird und mit der die Schätzwerte $\breve{\lambda}_{D,k,\ell}$ und $\hat{\lambda}_{D,k,\ell}$ verglichen werden, wird die geglättete Version des Spektrogramms der vorliegenden Störsignale verwendet:

$$\lambda_{D,k,\ell} = 0.95 \cdot \lambda_{D,k,\ell-1} + 0.05 \cdot |D_{k,\ell}|^2. \tag{7.22}$$

Um zu verdeutlichen, dass der MAP-basierte Postprozessor im Stande ist, das RLDS in den Zeit-Frequenz-Punkten mit Sprachaktivität besser als das IMCRA-Verfahren zu verfolgen, sind RLDS-Trajektorien in Abb. 7.3 (a) abgebildet. Hier sind der zeitliche Verlauf der Referenz $\lambda_{D,k,\ell}$ und die Verläufe von $\breve{\lambda}_{D,k,\ell}$ und $\hat{\lambda}_{D,k,\ell}$ aufgetragen, die jeweils vom IMCRA-Verfahren und vom MAPB-Postprozessor berechnet werden, und zwar für das *babble* Rauschen bei einem SNR von $0\,\text{dB}$ im Frequenzband $k = 97$, der mit der Frequenz von etwa $3\,\text{kHz}$ einhergeht. Im Unterschied zum IMCRA-Verfahren aktualisiert der MAPB-Postprozessor seine Schätzwerte auch während der Sprachaktivität und kann dadurch der

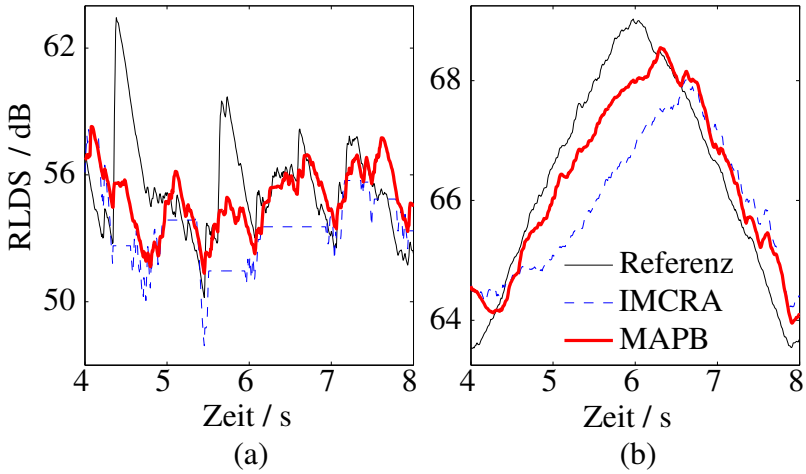


Abbildung 7.3.: RLDS-Trajektorien bei einem SNR von 0 dB: (a) gestört mit *babble* Rauschen im Frequenzband $k = 97$ (entspricht der Frequenz von etwa 3 kHz); (b) gestört mit NWNR gemittelt über alle Frequenzbänder.

Referenztrajektorie besser folgen. Die Zeit-Frequenz-Punkte mit Sprachaktivität lassen sich am Verlauf der IMCRA-Schätzwerte $\hat{\lambda}_{D,k,\ell}$ erkennen, denn während der Sprachaktivität hält das IMCRA-Verfahren seine Schätzwerte auf einem konstanten Wert. Die Abb. 7.3 (b) stellt die zeitlichen Verläufe der RLDS-Schätzwerte für einen Ausschnitt von NWNR beim SNR von 0 dB gemittelt über alle Frequenzbänder dar. Sie zeigt, dass der MAPB-Schätzer vor allem dem steigenden Rauschpegel besser folgen kann als das IMCRA-Verfahren.

Für eine quantitative Analyse der RLDS-Schätzer werden die Bewertungsmaße LEM aus (2.43) und LEV aus (2.45) leicht modifiziert. Zunächst wird ein logarithmischer Schätzfehler wie in (2.44), jedoch in Abhängigkeit von einem Verschiebungsindex i berechnet

$$\Delta_{k,\ell}(i) = \left| 10 \log_{10} \frac{\lambda^2_{D,k,\ell-i}}{\tilde{\lambda}_{D,k,\ell}} \right|, \tag{7.23}$$

wobei K die Anzahl der Frequenzbänder und L die Anzahl der Rahmen im vorliegenden Signal sind. Dabei werden in (7.23) für $\tilde{\lambda}_{D,k,\ell}$ entweder $\breve{\lambda}_{D,k,\ell}$ oder $\hat{\lambda}_{D,k,\ell}$ eingesetzt. Für jeden Verschiebungsindex i wird dann das LEM-Maß aus (2.43) berechnet

$$\mathrm{LEM}(i) = \frac{1}{K \cdot L} \sum_{k=1}^{K} \sum_{\ell=1}^{L} \Delta_{k,\ell}(i), \tag{7.24}$$

aus dem das *minimum* LEM (MLEM) berechnet wird

$$\mathrm{MLEM} = \min_{i} \left[\mathrm{LEM}(i) \right]. \tag{7.25}$$

Das MLEM-Maß ist bis auf die Minimumsuche über i dem LEM-Maß aus (2.43) ähnlich. Im Unterschied zum LEM-Maß wird der logarithmische Schätzfehler hier durch die gemeinsame zeitliche Verschiebung der RLDS-Schätzwerte um den Verschiebungsindex i so an die Referenz angepasst, dass das LEM-Maß für einen bestimmten Verschiebungsindex

$$i_{\min} = \operatorname*{argmin}_{i} \left[\mathrm{LEM}(i) \right] \tag{7.26}$$

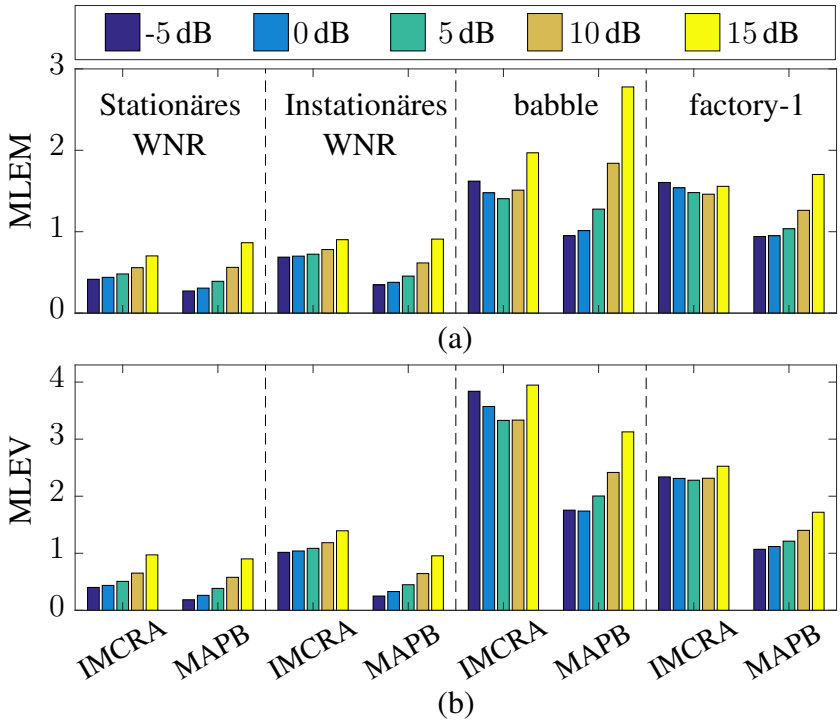


Abbildung 7.4.: Leistungsfähigkeit des IMCRA-Verfahrens und des MAPB-Postprozessors für verschiedene Rauschtypen und $\mathrm{SNR}_{\mathrm{IN}}$-Werte: (a) MLEM und (b) MLEV.

minimal wird. Diese Anpassung sorgt dafür, dass das MLEM-Bewertungsmaß keinen RLDS-Schätzer allein deswegen bevorzugt, dass er die zeitliche Verzögerung der Referenz nachahmt, die aufgrund der Rekursionsgleichung (7.22) zwangsläufig vorhanden ist. Das zweite Bewertungsmaß ist die mittlere Varianz des minimalen logarithmischen Schätzfehlers (engl. *minimum log-error variance*, MLEV), die das LEV-Maß aus (2.45) ist, wenn man in (2.46) und (2.47) $\Delta_{k,\ell}(i_{min})$ statt $\Delta(k,\ell)$ mit $K_s = 16$ verwendet.

In Abb. 7.4 wird die Leistungsfähigkeit des IMCRA-Verfahrens mit der des MAP-basierten Postprozessors verglichen, und zwar bezüglich der erreichten MLEM- und MLEV-Werte für unterschiedliche Rauschtypen und $\mathrm{SNR}_{\mathrm{IN}}$-Werte. Die dargestellten MLEM- und MLEV-Werte sind hier über die Signale mit weiblichen und männlichen Sprechern gemittelt. Der Abb. 7.4 (a) ist zu entnehmen, dass der MAPB-Postprozessor in vielen simulierten Umgebungen kleinere MLEM-Werte als das IMCRA-Verfahren liefert und somit den Verlauf der Referenz $\lambda_{D,k,\ell}$ im Mittel besser verfolgt. Nur bei den hohen $\mathrm{SNR}_{\mathrm{IN}}$-Werten von 15 dB und beim Rauschtyp *babble* mit 10 dB kann der MAPB-Postprozessor den IMCRA-Schätzer nicht übertreffen. Die resultierenden MLEV-Werte in Abb. 7.4 (b) offenbaren eine potenzielle Stärke des MAPB-Postprozessors und zwar eine kleine Streuung seiner Schätzwerte. Es fällt hier auf, dass der MAPB-Postprozessor in ausnahmslos allen simulierten Umgebungen deutlich kleinere MLEV-Werte als das IMCRA-Verfahren erreicht, was zu den besseren Eigenschaften der entstörten Signale führen kann.

Um Qualität der entstörten Ausgangssignale $\breve{s}(t)$ und $\hat{s}(t)$ quantitativ beurteilen zu können, werden außerdem die $\mathrm{MOS\text{-}LQO}_{\mathrm{WB}}$-Werte basierend auf dem vorhandenen ungestörten Sprachsignal $s(t)$ berechnet, die jeweils als $\mathrm{MOS\text{-}LQO}_{\mathrm{IMCRA}}$ und $\mathrm{MOS\text{-}LQO}_{\mathrm{MAPB}}$ bezeichnet werden. Da die Qualitätsunterschiede der Ausgangssignale der beiden Systemstufen von

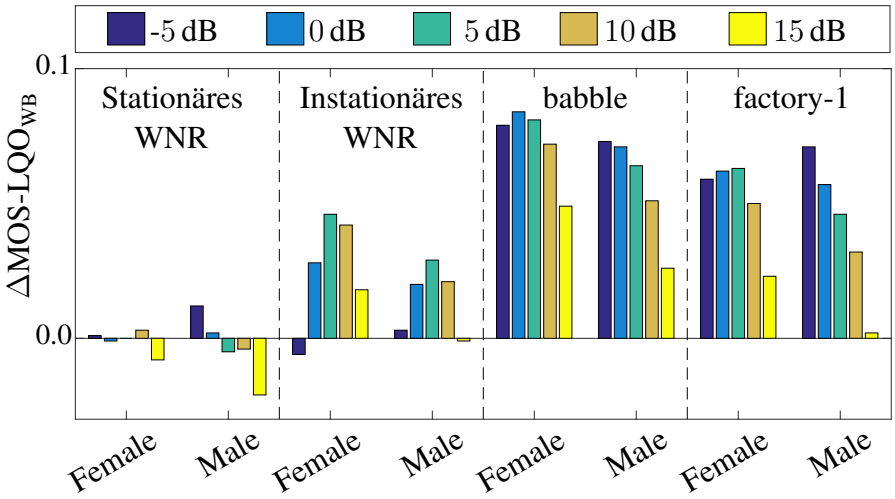


Abbildung 7.5.: Qualitätsunterschiede der entstörten Ausgangssignale $\Delta\text{MOS-LQO}_{\text{WB}}$ mit weiblichen und männlichen Sprechern und für verschiedene Rauschtypen und SNR_{IN}-Werte.

besonderem Interesse sind, wird anschließend die Differenz der MOS-LQO-Werte berechnet

$$\Delta\text{MOS-LQO}_{\text{WB}} = \text{MOS-LQO}_{\text{MAPB}} - \text{MOS-LQO}_{\text{IMCRA}}. \tag{7.27}$$

In Abb. 7.5 sind $\Delta\text{MOS-LQO}_{\text{WB}}$-Werte getrennt für Signale mit weiblichen und männlichen Sprechern dargestellt. Demnach belegen die Untersuchungen, dass bessere RLDS-Verfolgung einen positiven Effekt auf die Verbesserung der Qualität der entstörten Signale hat, insbesondere in Gegenwart von nichtstationären Störungen. Allerdings nimmt dieser positive Effekt mit steigenden SNR_{IN}-Werten etwas ab, was vermutlich an schlechter werdender RLDS-Schätzung gemessen in MLEM-Werten liegt.

Als Zwischenergebnis lässt sich festhalten, dass die durchgeführte Machbarkeitsstudie vielversprechende Erfolgsaussichten beim Einsatz des hergeleiteten MAPB-Postprozessors in einem System zur akustischen Signalentstörung in der Rolle eines RLDS-Schätzers aufzeigt. Ein Versuch, den Postprozessor für die Schätzung des LDS des ungestörten Sprachsignals $\lambda_{S,k,\ell}$ einzusetzen, blieb allerdings fruchtlos[5]. Dabei wurden Experimente mit kleineren Werten des Freiheitsgrades $\nu_0 < 40$ durchgeführt, um dem dynamischen Energieverlauf eines Sprachsignals folgen zu können. Dabei stellte sich heraus, dass der MAPB-Postprozessor in der aktuellen Ausführung noch nicht erwartungstreu ist und einen positiven Bias aufweist, der eine unzureichende RLDS-Schätzung bei hohen SNR_{IN}-Werten erklären kann und unbedingt korrigiert werden muss.

7.2. Qualitätsanalyse und Optimierung

Die Analyse der statistischen Eigenschaften des MAPB-Postprozessors aus Alg. 7.1 ist für seine optimale Leistungsfähigkeit von großer Bedeutung. Sie analytisch durchzuführen ist allerdings nur schwer möglich, denn der Postprozessor hängt von den Statistiken des LDS-Schätzers des ungestörten Sprachsignals $\breve{\lambda}_{S,k,\ell}$ ab, der seinerseits von statistischen Eigen-

[5]Prinzipiell könnte man Alg. 7.1 für die Schätzung von $\lambda_{S,k,\ell}$ beim gegebenen $\breve{\lambda}_{D,k,\ell}$ in der ersten Systemstufe in Abb. 7.2 verwenden.

schaften des in der ersten Systemstufe eingesetzten RLDS-Schätzers $\breve{\lambda}_{D,k,\ell}$ abhängig ist. Aus diesem Grund wird hier eine nummerische Qualitätsanalyse des MAPB-Postprozessors durchgeführt und zwar mit Hilfe der Monte-Carlo-Methode [Fis96]. Die Qualitätsanalyse deckt die drei folgenden Bereiche ab und wird in Unterabschnitt 7.2.1 beschrieben. Zunächst wird der Postprozessor hinsichtlich der Erwartungstreue und der Konsistenz untersucht. Danach wird seine Fähigkeit analysiert, ein nichtstationäres Rauschen zu verfolgen. Und zum Schluss wird seine Empfindlichkeit im Bezug auf die fehlerbehaftete Schätzung $\breve{\lambda}_{S,k,\ell}$ überprüft. Basierend auf der durchgeführten Qualitätsanalyse wird anschließend in Unterabschnitt 7.2.2 ein optimierter MAPB-basierter Postprozessor vorgestellt, der bessere statistische Eigenschaften aufweist als der Postprozessor aus Alg. 7.1. Dafür wird eine heuristische Optimierung durchgeführt, die zum einen eine fortlaufende Anpassung der Bandbreite des Postprozessors und zum anderen eine Biaskorrektur beinhaltet. Außerdem wird der Aufbau eines Systems zur spektralen Entstörung datailliert erläutert, in dessen zweiter Stufe der optimierte MAPB-Postprozessor als ein RLDS-Schätzer agiert. Die hier vorgestellten Qualitätsanalyse und Optimierung des MAP-basierten Postprozessors wurden zum ersten Mal in [CHU12] präsentiert. Einen Nachweis über die verbesserte Leistungsfähigkeit des optimierten Postprozessors wird in den experimentellen Untersuchungen erbracht, die in Abschnitt 7.3 beschrieben werden.

7.2.1. Nummerische Qualitätsanalyse des Postprozessors

Um die Qualitätsanalyse durchzuführen, wird eine Simulationsumgebung aufgebaut, deren Blockschaltbild in Abb. 7.6 dargestellt ist. Die Realisierungen des STFT-Spektrums eines gestörten Signals in einem Frequenzband werden als ein komplexwertiger mittelwertfreier weißer normalverteilter Zufallsprozess (WNZ) Y_ℓ modelliert, der aus einer additiven Überlagerung von WNZ S_ℓ und D_ℓ mit zeitveränderlichen Varianzen $\lambda_{S,\ell}$ und $\lambda_{D,\ell}$ entsteht mit $\ell \in [1;\, L]$, wobei L die Größe des Stichprobenumfangs ist. In jedem Zeitschritt ℓ berechnet der MAPB-Postprozessor entsprechend Alg. 7.1 einen Schätzwert für die Varianz des Rauschprozesses $\hat{\lambda}_{D,\ell}$ aus den Beobachtungen Y_ℓ und aus der Varianz $\tilde{\lambda}_{S,\ell}$, die je nach Position des Schalters S unterschiedlich gewählt wird. Das Betragsquadrat des momentane Wertes $|S_\ell|^2$ wird an den MAPB-Postprozessor weitergegeben, falls der Schalter auf den Knoten 1 gelegt wird. Ist der Schalter in Position 2, wird die wahre zeitveränderliche Varianz $\lambda_{S,\ell}$ dem

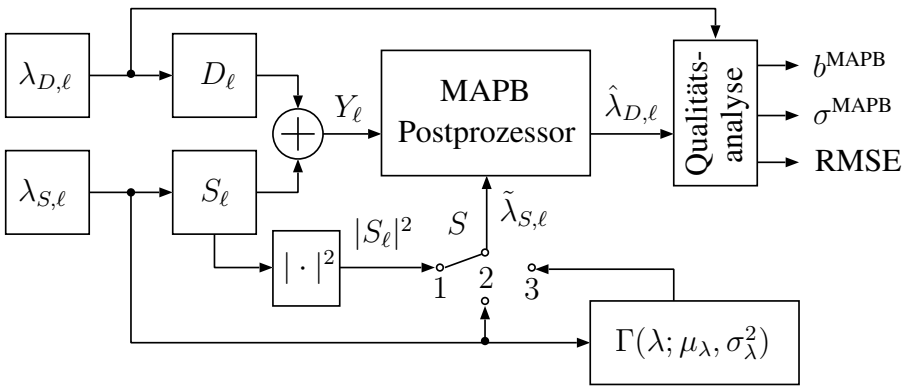


Abbildung 7.6.: Simulationsumgebung für Qualitätsanalyse des MAPB-Postprozessors.

MAPB-Postprozessor übermittelt. Sonst wird $\tilde{\lambda}_{S,\ell}$ aus einer Gammaverteilung gezogen

$$\tilde{\lambda}_{S,\ell} \sim \Gamma(\lambda; \mu_\lambda, \sigma_\lambda^2), \tag{7.28}$$

wobei der Mittelwert $\mu_\lambda = \lambda_{S,\ell}$ gleich dem wahren Wert $\lambda_{S,\ell}$ gesetzt wird und die Varianz $\sigma_\lambda^2 = var(\tilde{\lambda}_{S,\ell})$ verschiedene Werte annehmen darf. Im letzten Fall wird ein erwartungstreuer *a priori* SNR-Schätzer simuliert, dessen Unsicherheit variiert wird. Außerdem ist zu erwähnen, dass die Varianz des Zufallsprozesses S_ℓ in allen unseren Untersuchungen konstant $\lambda_{S,\ell} = \lambda_S$ ist und zwar entsprechend dem gewünschten Wert des *a priori* SNR ξ. Für jeden Wert von ξ werden K Simulationen durchgeführt, in denen die Schätzwerte $\hat{\lambda}_{D,k,\ell}$ mit $k \in [1,\ K]$ für unterschiedliche Werte von K berechnet werden. Um die Abhängigkeit der Schätzwerte $\hat{\lambda}_{D,k,\ell}$ vom Anfangswert zu reduzieren, wird $\hat{\lambda}_{D,k,0}$ in der Nähe des wahren Wertes gesetzt. Wie in Abb. 7.6 angedeutet, werden für die Qualitätsanalyse folgende Statistiken des MAPB-Postprozessors berechnet: ein durchschnittlicher relativer systematischer Schätzfehler (engl. *bias*) b^{MAPB}, eine mittlere Standardabweichung des Schätzfehlers σ^{MAPB} und die Wurzel des mittleren quadratischen Schätzfehlers RMSE. Alle diese Qualitätsmaße werden an den entsprechenden Stellen genauer definiert.

Im Unterabschnitt 7.1.1 haben wir zwei verschiedene Aktualisierungsmöglichkeiten des Freiheitsgrades ν_ℓ des MAPB-Postprozessors vorgestellt. Während die erste Variante, bei der der Freiheitsgrad anwachsen darf $\nu_{\ell+1} = \nu_\ell + 2$, ein stationäres Rauschen voraussetzt, wird die zweite Variante mit dem konstanten Freiheitsgrad $\nu_{\ell+1} = \nu_\ell = \nu_0$ für ein nichtstationäres Rauschen eingesetzt. Der Wert des Freiheitsgrades ν_ℓ bestimmt die Bandbreite des Postprozessors wie eines Tiefpassfilters, wobei große Werte von ν_ℓ mit einer schmalbandigen Filterfunktion (starke Glättung) und kleine mit der breitbandigen (schwache Glättung) einhergehen. Im Weiteren wird zunächst die Erwartungstreue und die Konsistenz des MAPB-Schätzers für beide Aktualisierungsmöglichkeiten untersucht, die als eine EK Analyse bezeichnet wird. Dabei wird der Zufallsprozess D_ℓ mit einer konstanten Varianz $\lambda_{D,\ell} = \lambda_D = 1$ generiert und der Schalter S auf die Position 1 gelegt, wie in Abb. 7.6 dargestellt ist. Ähnliche nummerische Untersuchung wird für den RLDS-Rauschschätzer in [Mar94] für den Fall der Abwesenheit des Sprachsignals durchgeführt und zwar mit dem Ziel den systematischen Schätzfehler des MS-Verfahrens zu bestimmen. Die Verfolgungsfähigkeit und die Sensitivitätsanalyse des Postprozessors werden nur für den konstanten Freiheitsgrad analysiert.

EK Analyse für den steigenden Freiheitsgrad: Im Fall einer stationären Störung wird der Freiheitsgrad mit der Aktualisierungsgleichung $\nu_{\ell+1} = \nu_\ell + 2$ berechnet. Die für diesen Fall resultierenden Verteilungen der Schätzwerte $\hat{\lambda}_{D,k,L}$, die für unterschiedliche Größen des Stichprobenumfangs $L = \{250, 1000, 2500, 5000, 10000\}$ und verschiedene Werte von ξ im Rahmen der $K = 10^4$ Simulationen berechnet werden, sind in Abb. 7.7 dargestellt. Aus den Schätzwerten $\hat{\lambda}_{D,k,\ell}$ lässt sich ein relativer Bias b_ℓ berechnen

$$b_\ell = \frac{\bar{\lambda}_{D,\ell} - \lambda_D}{\lambda_D} \quad \text{mit} \quad \bar{\lambda}_{D,\ell} = \frac{1}{K} \sum_{k=1}^{K} \hat{\lambda}_{D,k,\ell}. \tag{7.29}$$

Die Simulationsergebnisse vermitteln den Eindruck, dass der MAPB-Schätzer für die negativen Werte von ξ gemessen in dB asymptotisch erwartungstreu ist mit $\lim_{\ell\to\infty} b_\ell = 0$, siehe Abb. 7.7 (a). Außerdem sinkt in diesem Fall die Varianz der Schätzwerte $\hat{\lambda}_{D,k,\ell}$, was für die Konsistenz des MAPB-Schätzers spricht.

Allerdings geht die Erwartungstreue des Schätzers in der Gegenwart einer stärkeren Störung durch den Zufallsprozess S_ℓ verloren, und ein positiver Bias wird offenbar, der mit

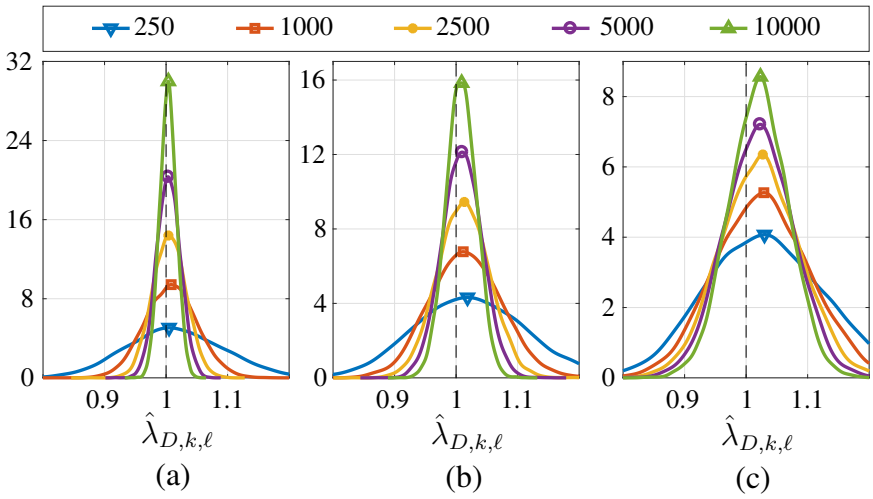


Abbildung 7.7.: Histogramme der MAPB-Schätzwerte $\hat{\lambda}_{D,k,L}$ berechnet mit Alg. 7.1 für verschiedene Größen des Stichprobenumfangs $L = \{250, 1000, 2500, 5000, 10000\}$ bei unterschiedlichen Werten des *a priori* SNR ξ (a) -5 dB, (b) 0 dB und (c) 5 dB.

steigenden Werten von ξ weiter wächst, wie in den Abb. 7.7 (b) und (c) zu sehen ist. Zudem fällt hier auf, dass auch die Konsistenz des Schätzers unter dem Verlust der Erwartungstreue leidet. Das systematische Überschätzen der wahren Varianz λ_D und mangelnde Konsistenz der MAPB-Schätzers erklären die mangelnde Leistungsfähigkeit des Postprozessors in Unterabschnitt 7.1.2. Bevor jedoch die Maßnahmen zur Biaskorrektur ergriffen werden können, muss das Biasverhalten des Schätzers auch beim konstanten Freiheitsgrad untersucht werden.

EK Analyse für den konstanten Freiheitsgrad: Ist eine Störung nichtstationär, wird der Freiheitsgrad konstant $\nu_{\ell+1} = \nu_\ell = \nu_0$ gehalten. Die Parametrisierung des MAPB-Postprozessors in diesem Fall ist für eine Verfolgung der nichtstationären spektralen Rauschleistungsdichte erforderlich und stellt sich als eine praxisrelevante Implementierung des MAPB-Schätzers heraus, wie die Machbarkeitsstudie in Unterabschnitt 7.1.2 zeigte. Um das Biasverhalten des Schätzers in dieser Implementierung möglichst ohne den Einfluss seiner Anlaufphase zu erfassen, wird ein durchschnittlicher relativer Bias b^{MAPB} basierend auf den letzten $L/2$ Werten b_ℓ für $\ell \in [L/2+1\,;L]$ mit $L = 10^4$ berechnet

$$b^{\text{MAPB}} = \frac{1}{L/2} \sum_{\ell=L/2+1}^{L} b_\ell \tag{7.30}$$

und zwar für unterschiedliche Werte des Freiheitsgrades ν_0 und verschiedenen Werte von ξ. Außerdem wird die durchschnittliche Standardabweichung des MAPB-Schätzers bestimmt

$$\sigma^{\text{MAPB}} = \frac{1}{L/2} \sum_{\ell=L/2+1}^{L} \sqrt{\frac{1}{K-1} \sum_{k=1}^{K} \left(\hat{\lambda}_{D,k,\ell} - \bar{\lambda}_{D,\ell}\right)^2}. \tag{7.31}$$

Die resultierenden Werte von b^{MAPB} und σ^{MAP} sind in Abb. 7.8 als Funktion des *a priori* SNR ξ für $K = 10^3$ dargestellt. In Abb. 7.8 (a) ist es deutlich zu sehen, dass b^{MAPB} immer positiv ist. Dabei wächst der Bias mit steigenden Werten von ξ an, wie dies auch beim

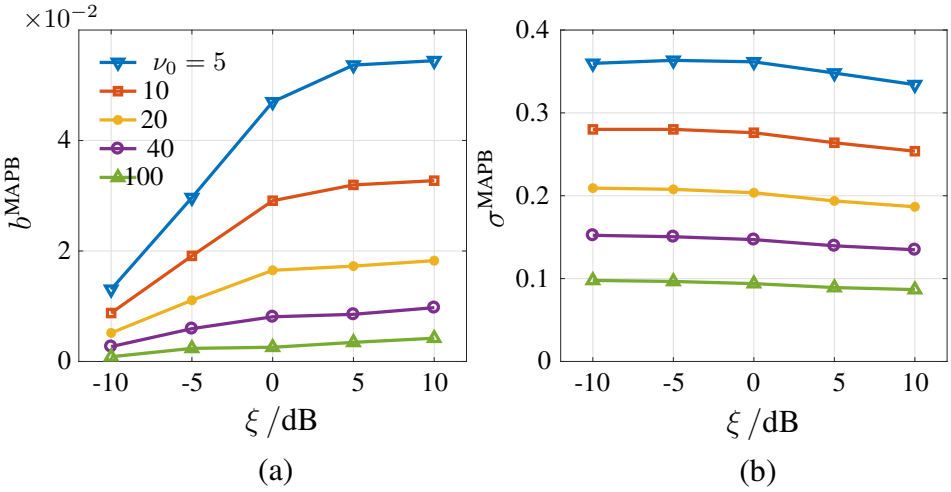


Abbildung 7.8.: (a) Durchschnittlicher relativer Bias b^{MAPB} und (b) durchschnittliche Standartabweichung σ^{MAPB} des MAPB-Postprozessors als Funktion des *a priori* SNR ξ für verschiedene Freiheitsgrade $\nu_0 = \{5, 10, 20, 40, 100\}$.

steigenden Freiheitsgrad in Abb. 7.7 der Fall war. Allerdings, wie Abb. 7.8 (b) zeigt, hängt die Streuung der Schätzwerte $\hat{\lambda}_{D,k,\ell}$ kaum von ξ ab, im Unterschied zu den Ergebnissen aus Abb. 7.7. Dies ist eine weitere positive Eigenschaft des MAPB-Schätzers für den konstanten Freiheitsgrad. Ansonsten fällt auf, dass sowohl b^{MAPB} als auch λ^{MAP} sehr stark von den Werten von ν_0 abhängen. Um also den mittleren quadratischen Fehler des MAPB-Schätzers möglichst zu reduzieren, soll ν_0 so groß wie möglich gewählt werden. Es sei denn, dass dadurch die Eigenschaft des Schätzers verschlechtert wird, dem zeitvarianten RLDS bei einer nichtstationären Störung zu folgen. Ein optimaler Kompromiss zwischen diesen beiden Anforderungen soll im nächsten Abschnitt zur sinnvollen Wahl des Freiheitsgrades ν_0 führen.

Analyse der Verfolgungsfähigkeit: Um die Eigenschaften des MAPB-Schätzers im Bezug auf das Verfolgen eines zeitvarianten RLDS zu untersuchen, wird zum einen der Schalter S in Abb. 7.6 auf die Position 2 gesetzt, sodass es $\tilde{\lambda}_{S,\ell} = \lambda_{S,\ell}$ gilt, und zum anderen ein Zufallsprozess D_ℓ mit der zeitveränderlichen Varianz simuliert

$$\lambda_{D,\ell} = 1 + \sin^2\left(\frac{2\pi \cdot \ell}{L}\right), \tag{7.32}$$

wobei $L = 10^4$ gewählt wird. In den $K = 10^3$ Experimenten werden dann die MAPB-Schätzwerte $\hat{\lambda}_{D,k,\ell}$ für unterschiedliche Werte von ν_0 und ξ berechnet. Anschließend wird aus der Referenz $\lambda_{D,\ell}$ und den vorliegenden Schätzwerten $\hat{\lambda}_{D,k,\ell}$ der RMSE-Wert bestimmt

$$\text{RMSE} = \frac{1}{L - l_0} \sum_{\ell=l_0+1}^{L} \sqrt{\frac{1}{K} \sum_{k=1}^{K} \left(\hat{\lambda}_{D,k,\ell} - \lambda_{D,\ell}\right)^2}. \tag{7.33}$$

Und wieder, um den Einfluss der Anlaufphase des MAPB-Schätzers zu reduzieren und sich nur auf die gewünschte Verfolgungsfähigkeit des MAPB-Postprozessors zu konzentrieren, werden für die Berechnung der RMSE-Werte nur die Schätzwerte ab dem Index $l_0 = 50$ verwendet. Die resultierenden RMSE-Werte sind in Abb. 7.9 (a) als Funktion von ν_0 für verschiedene Werte des *a priori* SNR ξ dargestellt. Wie erwartet, verschlechtert sich die Leistungsfähigkeit des MAPB-Schätzers für größer werdende Werte von ξ. Außerdem wachsen

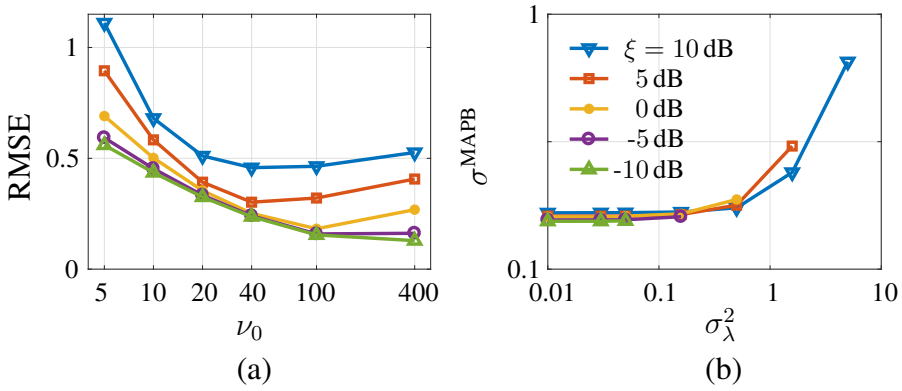


Abbildung 7.9.: Verfolgungsfähigkeit und Sensitivitätsanalyse des MAPB-Schätzers für verschiedene Werte des *a priori* SNR ξ: (a) Verfolgungsfähigkeit gemessen in RMSE-Werten als Funktion von ν_0 und (b) Sensitivität der RLDS-Schätzwerte $\hat{\lambda}_{D,\ell}$ gemessen mit σ^{MAPB} aus (7.31) als Funktion von σ_λ^2 aus (7.28) für $\nu_0 = 40$.

die RMSE-Werte nicht nur für kleine Werte von ν_0 aufgrund des steigenden Bias, sondern auch für große Werte von ν_0 aufgrund einer starken Glättung, welche die Verfolgungseigenschaften des Schätzers einschränkt. Eine gute Nachricht ist allerdings, dass es einen optimalen Wert für den Parameter ν_0 gibt, der jedoch eine Abhängigkeit vom *a priori* SNR-Wert ξ aufweist. Weitere Untersuchungen zeigten, dass ein konstanter Freiheitsgrad von $\nu_0 = 40$, der sich für $\xi > 0$ dB laut Abb. 7.9 (a) als Optimum erweist, eine gute Wahl ist.

Sensitivität gegenüber einer fehlerhaften Schätzung von $\lambda_{S,\ell}$: Zum Schluss der Qualitätsanalyse soll noch die Sensitivität des MAPB-Schätzers gegenüber der fehlerhaften Schätzung von $\lambda_{S,\ell}$ untersucht werden. Dafür wird $\tilde{\lambda}_{S,\ell}$ aus einer Gammaverteilung (7.28) mit dem Mittelwert $\mu_\lambda = \lambda_{S,\ell}$ und der Varianz $\sigma_\lambda^2 = var(\tilde{\lambda}_{S,\ell})$ für unterschiedliche Werte von σ_λ^2 gezogen, wofür der Schalter S an die Position 3 gelegt wird, siehe Abb. 7.6. Im Rahmen dieser Analyse wird eine konstante Varianz $\lambda_{D,\ell} = \lambda_D = 1$ und der konstante Freiheitsgrad $\nu_0 = 40$ verwendet. Für jeden Wert des *a priori* SNR ξ werden $K = 10^3$ Simulationen durchgeführt. Die resultierenden Werte für die Standardabweichung des MAPB-Schätzers σ^{MAPB} sind als Funktion der Varianz der Gamma-Verteilung σ_λ^2 in Abb. 7.9 (b) für unterschiedliche Werte von ξ dargestellt. Demnach bleibt der MAPB-Schätzer gegenüber der Schätzfehlervarianz von λ_S im relativ großen Wertebereich von σ_λ^2 robust. Allerdings wächst seine Schätzfehlervarianz ab einem bestimmten Wert von σ_λ^2 stark. Ansonsten zeigen die Simulationen wieder, dass die Varianz des Schätzers für einen konstanten Wert von ν_0 relativ unabhängig vom Wert des *a priori* SNR ξ ist.

Bevor die heuristische Optimierung des MAPB-Postprozessors vorgestellt wird, soll erwähnt werden, dass die aufeinander folgenden Realisierungen des Zufallsprozesses D_ℓ in allen bisherigen Untersuchungen als unkorreliert $\text{cov}(D_\ell, D_{\ell-i}) = 0$ für $\forall i \neq \ell$ angenommen wurden. Dasselbe galt auch für den Zufallsprozess S_ℓ. In einem System zur Sprachsignalentstörung wird diese Annahme durch die Überlappung der aufeinander folgenden Rahmen bei der STFT verletzt und zwar auch für das weiße Rauschen. Die Autokorrelationsfunktion $\phi_{D,D}(i)$ hängt dann von solchen STFT-Parametern ab, wie das Analysefenster und der Blockvorschub [Mar94]. Selbstverständlich verändern sich die Eigenschaften des MAPB-Schätzers, wenn er auf korrelierten Beobachtungen arbeitet. Aus diesem Grund soll bei der Biaskorrektur auch der Fall der zeitlich korrelierten Signalen betrachtet werden.

7.2.2. Bandbreitenanpassung und Biaskorrektur

Die Qualitätsanalyse des MAPB-Postprozessors offenbarte sowohl Stärken als auch Schwächen des in Unterabschnitt 7.1.1 entwickelten Rauschschätzers. Die bevorstehende Optimierung soll die Stärken aufrechterhalten und die Schwächen mildern ohne dabei die Parameteranzahl des Postprozessors unnötig zu vergrößern. Während die experimentellen Untersuchungen in Unterabschnitt 7.1.2 die mangelnde Leistungsfähigkeit des MAPB-Postprozessors beim hohen $\mathrm{SNR}_{\mathrm{IN}}$ aufzeigten, lieferte die Qualitätsanalyse den Hauptgrund für dieses Verhalten, und zwar einen $\mathrm{SNR}_{\mathrm{IN}}$-abhängigen Bias. Dieser soll im Folgenden mit Hilfe von zwei Optimierungsschritten reduziert werden. Und da die beiden Schritte auf den Ergebnissen der nummerischen Analyse beruhen, ist hier die Rede von einer heuristischen Optimierung. Nachdem im ersten Optimierungsschritt die Anpassung der Bandbreite des Postprozessors durchgeführt wird, soll die anschließende Biaskorrektur die Eigenschaften des MAPB-Rauschschätzers bezüglich seiner Erwartungstreue verbessern.

Bandbreitenanpassung: In Abb. 7.8 (a) wurde ersichtlich, dass es bei hohen Werten des *a priori* SNR vorteilhaft ist, den Freiheitsgrad des Postprozessors ν_0 kontrollierbar zu erhöhen, um keinen unnötig großen Bias zuzulassen. Wie bereits erwähnt, verringert die Vergrößerung von ν_0 die Bandbreite des Postprozessors und somit seine Fähigkeit, dem nichtstationären Rauschen zu folgen. Ein solches Verhalten ist besonders in der Anwesenheit des Sprachsignals mit starker Leistung durchaus wünschenswert, wie weitere experimentelle Untersuchungen gezeigt haben. Eine Reduzierung der Verfolgungsfähigkeit in Sprachsignalanwesenheit ist auch bei solchen modernen Rauschschätzern zu finden wie MCRA in [CB01a] und IMCRA-Verfahren in [Coh03], in denen die Konstante α der rekursiven Glättung (7.9) in Anwesenheit des Sprachsignals größere Werte annehmen darf. Somit wird der Schätzwert der Rauschleistungsdichte in solchen Zeit-Frequenz-Punkten kaum verändert. In den Zeit-Frequenz-Punkten, in denen die spektrale Leistungsdichte des Sprachsignals besonders groß ist, wird die Glättungskonstante sogar auf den Wert $\alpha = 1$ gesetzt, was einem konstanten RLDS-Schätzwert entspricht. Ähnliche Strategie wird auch im VAD-RA-Rauschschätzer verwendet [Hir93, HE95].

Eine Möglichkeit, die Bandbreite des Postprozessors zu verändern, ist die Steuerung des Freiheitsgrades ν_0 in Abhängigkeit vom momentanen Wert des *a priori* SNR $\xi_{k,\ell}$ mit

$$\nu_{k,\ell} = \nu_0 + \frac{\Delta\nu}{\pi} \cdot \arctan\left(\xi_{k,\ell}\right), \tag{7.34}$$

wobei $\Delta\nu \in (0; 2\nu_0)$ ein zusätzlicher Parameter ist, der im Weiteren als Regelbereich des Freiheitsgrades bezeichnet wird. Somit wird der Freiheitsgrad $\nu_{k,\ell}$ zeitvariant und frequenzabhängig mit möglichen Werten im Bereich $\nu_0 - \frac{\Delta\nu}{2} \leq \nu_{k,\ell} \leq \nu_0 + \frac{\Delta\nu}{2}$. Man beachte, dass der Regelbereich $\Delta\nu$ nicht allzu groß gewählt werden darf, denn für $\nu_{k,\ell} = \nu_0 - \frac{\Delta\nu}{2}$ steigt der Bias, und für $\nu_{k,\ell} = \nu_0 + \frac{\Delta\nu}{2}$ wird die Verfolgungsfähigkeit des Postprozessors beeinträchtigt. Aus den Ergebnissen der Analyse der Verfolgungsfähigkeit des Postprozessors wird der Freiheitsgrad $\nu_0 = 40$ gesetzt. In weiteren Experimenten hat sich der Wert des Regelbereichs $\Delta\nu = 10$ als gut erwiesen.

Biaskorrektur: Die vorgeschlagene Bandbreitenanpassung muss bei der Bestimmung der Biaskorrektur unbedingt berücksichtigt werden. Außerdem soll der Bias des MAPB-Postprozessors in einer möglichst praxisnahen Umgebung mit korrelierten Beobachtungen betrachtet werden und zwar mit den festgelegten Parametern $\nu_0 = 40$ und $\Delta\nu = 10$. Man

beachte, dass der Bias eines Schätzers von den Korrelationseigenschaften der Beobachtungen abhängig sein kann, siehe Untersuchungen in [Mar94, VM06]. Um stationäre korrelierte Rauschsignale $\tilde{D}_\ell$ mit der Varianz $\lambda_D = 1$ zu erzeugen, wird das weiße Rauschen der NOISEX-92-Datenbank verwendet. Dieses wird mit der STFT in den Frequenzbereich transformiert und zwar mit folgenden Parametern: die FFT-Länge von 2^{12} Abtastwerten, der Fenstervorschub 2^{10} Abtastwerte und das Hamming-Fenster als Analyse-Fenster. Anschließend werden die STFT-Koeffizienten in jedem Frequenzband so normiert, dass sie die Varianz $\lambda_D = 1$ aufweisen und als $\tilde{D}_\ell$ verwendet werden. Man beachte, dass $\tilde{D}_\ell$ im Unterschied zu D_ℓ in Abb. 7.6 nicht mehr weiß ist. Wie die Qualitätsanalyse bzgl. der Erwartungstreue wird auch für Simulationen mit korrelierten Signalen immer noch die Simulationsumgebung aus Abb. 7.6 mit der Schalterposition $S = 1$ verwendet. Verläufe der resultierenden Biaswerte b^{MAPB} des MAPB-Postprozessors mit $\nu_0 = 40$ und $\Delta\nu = 10$ für unkorrelierte und korrelierte Signale bei unterschiedlichen Werten von ξ sind in Abb. 7.10 (a) für $K = 2000$ und $L = 2500$ dargestellt. Ähnlich wie in Abb. 7.8 steigt der Bias mit den wachsenden *a priori* SNR-Werten sowohl bei unkorrelierten als auch bei korrelierten Signalen. Der auf den korrelierten Signalen berechneter Bias ist dabei stets positiv und etwas größer als bei Verarbeitung der unkorrelierten Signale. Der Bereich der Werte des *a priori* SNR, für die der Bias berechnet wird, ist in Abb. 7.10 (b) motiviert. Hier ist ein Histogramm der wahren *a priori* SNR-Werte $\xi_{k,\ell}$ gemessen in dB für ein gestörtes Sprachsignal dargestellt, das ein globales SNR_{IN} von 5 dB aufweist. Dabei weist das verwendete Sprachsignal ein mittleres *a priori* SNR-Wert von etwa 17 dB auf, was wegen der Darstellung des *a priori* SNR-Werte in Dezibel aus dem Bild nicht erkannt werden kann. Viele kleine *a priori* SNR-Werte sind der Dünnbesetztheit der Sprachsignale zu verdanken.

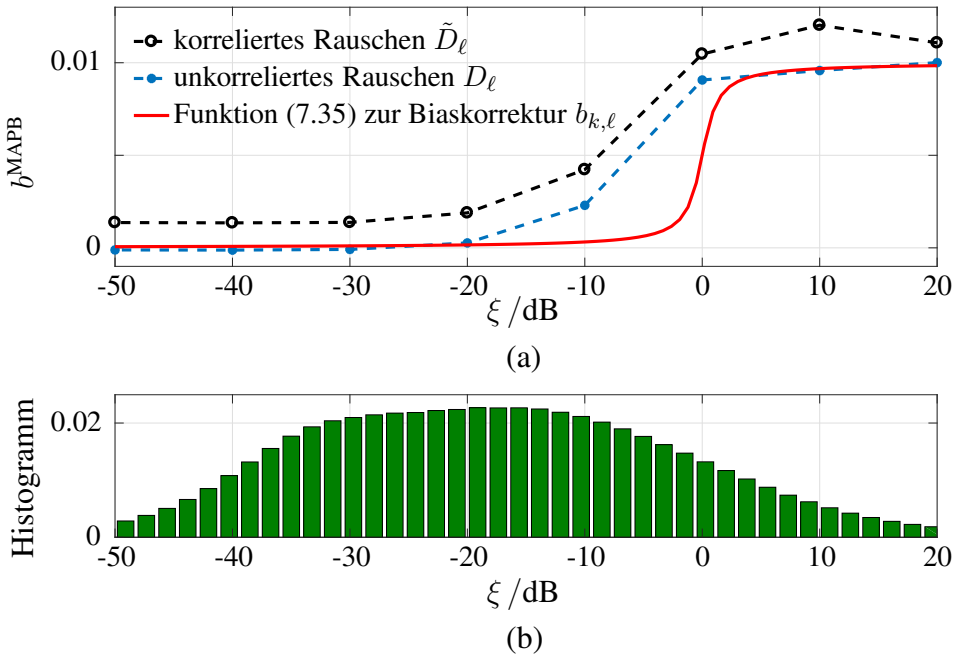


Abbildung 7.10.: Bias des MAPB-Postprozessors als Funktion des mittleren *a priori* SNR ξ: (a) MAPB-Bias für $\nu_0 = 40$ und $\Delta\nu = 10$ mit der vorgeschlagenen Funktion zur Biaskorrektur $b_{k,\ell}$ aus (7.35), (b) Verteilung der wahren *a priori* SNR-Werte $\xi_{k,\ell}$ eines gestörten Sprachsignals mit globalem SNR_{IN} von 5 dB.

Um einen MAPB-Rauschschätzer mit einem reduzierten Bias zu erhalten, wird der Bias des MAPB-Postprozessors mit folgender Funktion approximiert

$$b_{k,\ell} = \beta_{\max} \cdot \left(\frac{\arctan(\xi_{k,\ell})}{\pi} + \frac{1}{2} \right), \tag{7.35}$$

die einen einzigen Parameter den Biaskompensationsfaktor $\beta_{\max} = 0.01$ besitzt. Und obwohl die Biasverläufe aus Abb. 7.10 mit Hilfe zusätzlicher Parameter sicherlich besser approximiert werden könnten, wird die Funktion (7.35) aufgrund ihrer einfachen Parametrisierung bevorzugt. Die Verwendung dieser konservativen Funktion für die Biaskompensation verhindert außerdem eine Überschätzung des RLDS, was für eine gute Qualität der prozessierten Sprachsignale sehr wichtig ist. Man beachte, dass durch eine alternative Wahl von $\beta_{\max}$ die Sprunghöhe von (7.35) bei Bedarf angepasst werden kann. Mit (7.35) kann der optimierte Schätzwert des MAPB-Postprozessors mit reduziertem Bias wie folgt berechnet werden

$$\hat{\lambda}_{D,k,\ell} = (1 - b_{k,\ell}) \cdot \hat{\lambda}_{k,\ell}, \tag{7.36}$$

wobei $\hat{\lambda}_{k,\ell}$ die nummerisch berechnete (biasbehaftete) Nullstelle der Funktion $g(\lambda)$ aus (7.17) im (k, ℓ)-ten Zeit-Frequenz-Punkt ist.

Zusammenfassend lässt sich festhalten, dass die beiden Optimierungsschritte, die Bandbreitenanpassung und die Biaskorrektur, zwei zusätzliche Parameter mit sich mitbringen. Im Vergleich zum MAPB-Postprozessor in Alg. 7.1 hat der optimierte Postprozessor insgesamt drei Parameter: den Freiheitsgrad ν_0, den Regelbereich des Freiheitsgrades $\Delta\nu$ und den Biaskompensationsfaktor $\beta_{\max}$. Im Vergleich zu einigen anderen modernen RLDS-Schätzern

Algorithm 7.2 Der optimierte MAP-basierte RLDS-Schätzer

Input: Spektrogramm $|y_{k,\ell}|^2$, Schätzwerte $\breve{\lambda}_{S,k,\ell}$ und *a priori* SNR $\breve{\zeta}_{k,\ell}$
Parameter: Freiheitsgrad ν_0, sein Regelbereich $\Delta\nu$ und Biaskompensationsfaktor $\beta_{\max}$
Output: LDS-Schätzwert des Störsignals $\hat{\lambda}_{D,k,\ell}$

loop
  Für jeden Frequenzband $k \in [1;\, K]$ berechne
  for $\ell = 1$ **to** L **do**
    if $\ell == 1$ **then**
      den initialen LDS-Schätzwert $\hat{\lambda}_{D,k,1} = \frac{1}{2} \cdot |y_{k,1}|^2$
    else
      den Freiheitsgrad $\nu_{k,\ell} = \nu_0 + \frac{\Delta\nu}{\pi} \cdot \arctan\left(\breve{\zeta}_{k,\ell}\right)$ über die Bandbreitenanpassung
      die markanten Punkte $\lambda_a = \frac{\nu_0}{\nu_0+2} \cdot \lambda_{k,\ell-1}$ und $\lambda_b = |y_{k,\ell}|^2 - \breve{\lambda}_{S,k,\ell} > 0$
      die Nullstelle $\hat{\lambda}_{k,\ell}$ von $g_{k,\ell}(\lambda) = (\nu_0 + 2)(\lambda - \lambda_a)(\breve{\lambda}_{S,k,\ell} + \lambda)^2 + 2(\lambda - \lambda_b)\lambda^2$ in Intervallgrenzen $\hat{\lambda}_{k,\ell} \in (\lambda_{\mathrm{U}};\, \lambda_{\mathrm{O}})$ mit $\lambda_{\mathrm{U}} = \min(\lambda_a, \lambda_b)$ und $\lambda_{\mathrm{O}} = \max(\lambda_a, \lambda_b)$
      den Bias $b_{k,\ell} = \beta_{\max} \cdot \left(\frac{1}{\pi} \arctan(\breve{\zeta}_{k,\ell}) + \frac{1}{2}\right)$ für die Biaskorrektur
      den optimierten LDS-Schätzwert $\hat{\lambda}_{D,k,\ell} = (1 - b_{k,\ell}) \cdot \hat{\lambda}_{k,\ell}$ mit reduziertem Bias
    end if
    den Skalierungsparameter $\lambda_{k,\ell} = \frac{\nu_0+2}{\nu_0} \cdot \hat{\lambda}_{D,k,\ell}$
  end for
end loop

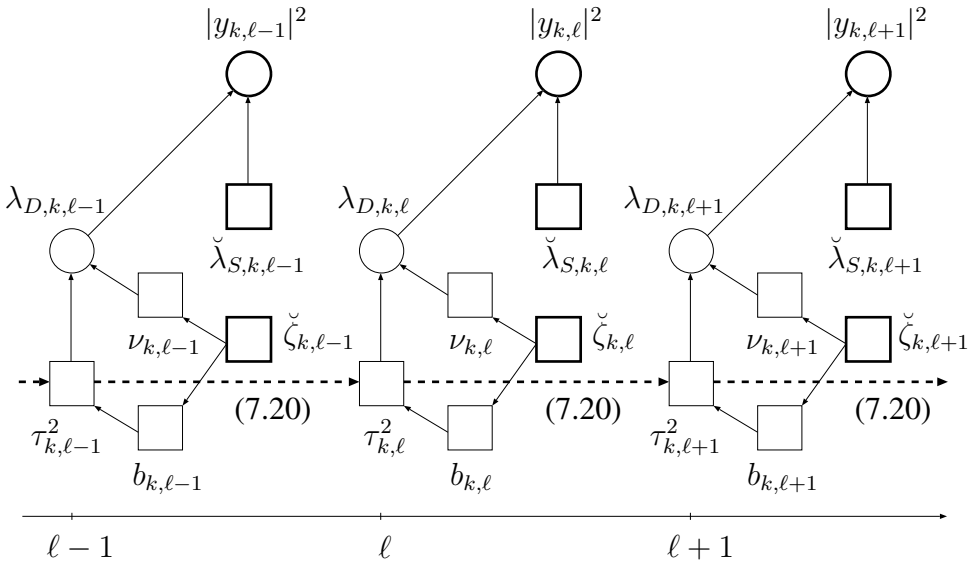


Abbildung 7.11.: Graphische Darstellung der statistischen Zusammenhänge des optimierten MAPB-Postprozessors im k-ten Frequenzband.

wie [CB01a] und [Coh03] ist die Anzahl der Parameter des MAPB-Postprozessors immer noch relativ gering. Außerdem benötigt der optimierte Postprozessor eine zusätzliche Eingangsgröße – das *a priori* SNR $\xi_{k,\ell}$. Die experimentellen Untersuchungen haben gezeigt, dass es vorteilhaft ist, die geglättete Version des *a priori* SNR im Postprozessor zu verwenden, die im OMLSA-Verfahren der ersten Systemstufe aus den Schätzwerten des DD-Verfahrens ohnehin berechnet wird und zwar mit Hilfe der Rekursionsgleichung

$$\breve{\zeta}_{k,\ell} = \alpha_\zeta \cdot \breve{\zeta}_{k,\ell-1} + (1-\alpha_\zeta) \cdot \breve{\xi}_{k,\ell} \tag{7.37}$$

mit der Glättungskonstante $\alpha_\zeta = 0{,}7$ wie in [CB01b]. Somit lassen sich die Berechnungsschritte des optimierten MAPB-Postprozessors in Alg. 7.2 zusammenfassen. Die statistischen Zusammenhänge des optimierten MAP-basierten Postprozessors im k-ten Frequenzband sind graphisch in Abb. 7.11 dargestellt.

Durch die Verwendung des geglätteten *a priori* SNR $\breve{\zeta}_{k,\ell}$ für die Bandbreitenanpassung und die Biaskorrektur wird der optimierte MAPB-Postprozessor stärker mit der ersten Stufe des Systems zu akustischen Signalentstörung aus Abb. 7.2 gekoppelt als seine nicht optimierte Version aus Alg. 7.1. Während der nicht optimierte MAPB-Postprozessor nur von den Schätzergebnissen der RLDS und von dem *a priori* SNR-Schätzer abhängig war, benötigt der optimierte Postprozessor außerdem noch $\breve{\zeta}_{k,\ell}$, das im OMLSA-Verfahren berechnet wird[6]. Wie weitere Experimente gezeigt haben, lohnt es sich, den Gedanken der noch engeren Kopplung der beiden Systemstufen aus Abb. 7.2 weiter zu verfolgen, denn dies führt zur recheneffizienteren Realisierung der zweiten Systemstufe. Die erwähnte Kopplung der beiden Systeme betrifft das OMLSA-Verfahren aus [CB01b], das für die Berechnung seiner Filterfunktion $G_{k,\ell}$ die Wahrscheinlichkeit der Sprachsignalabwesenheit $q_{k,\ell}$ schätzt. Wie auch beim nicht optimierten MAP-basierten RLDS-Schätzer zeigten die Untersuchungen auch hier, dass die Schätzwerte von $\breve{q}_{k,\ell}$ der ersten Systemstufe für die Signalentstörung bereits gut

[6]Falls für die Signalentstörung eine andere als die OMLSA-Filterfunktion verwendet wird, soll das geglättete *a priori* SNR entsprechend (7.37) zusätzlich ausgerechnet werden.

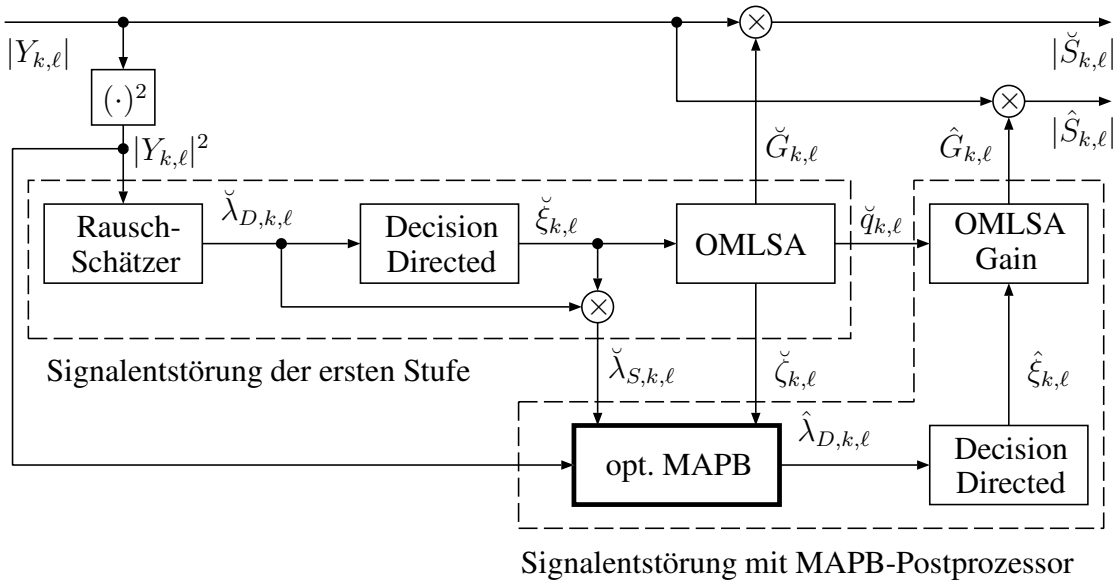


Abbildung 7.12.: Zweistufiges System zur Entstörung spektraler Amplituden unter Verwendung eines beliebigen RLDS-Schätzers und des optimierten MAPB-Postprozessors.

genug sind, und sie müssen in der zweiten Stufe nicht noch einmal berechnet werden. Somit darf die spektrale Filterfunktion des OMLSA-Verfahrens in der zweiten Stufe die Schätzwerte $\breve{q}_{k,\ell}$ verwenden. Das recheneffiziente und eng gekoppelte Gesamtsystem zur Signalentstörung mit dem optimierten MAPB-Postprozessor ist in Abb. 7.12 dargestellt. Man beachte, dass in der ersten Systemstufe ein beliebiger RLDS-Schätzer statt des IMCRA-Verfahrens verwendet werden kann. Ansonsten wird für die Berechnung der entstörten Zeitsignale der beiden Systemstufen $\breve{s}(t)$ und $\hat{s}(t)$ die Phase des gestörten Spektrums $Y_{k,\ell}$ verwendet, wie in Abb. 2.1 angedeutet.

Der entwickelte MAPB-Postprozessor ist für eine sequentielle Signalverarbeitung konzipiert, d. h., dass die RLDS-Schätzwerte des ℓ-ten Blocks auf Basis aller Werte des Leistungsdichtespektrums des gestörten Signals berechnet werden, die bis zum ℓ-ten Block einschließlich im Signal vorliegen. Mit anderen Worten, das gestörte Signal muss nicht in seiner gesamten Länge für eine Signalverarbeitung vorliegen. In diesem Fall wird in dieser Arbeit von einer *online* Signalverarbeitung gesprochen. Somit ermöglicht die zuletzt eingeführte Optimierung die eventuelle Parallelisierung der Signalverarbeitung in einem System mit fortlaufend eintreffenden Daten. Während in der ersten Systemstufe die Wahrscheinlichkeit der Sprachabwesenheit berechnet wird, können in der zweiten Stufe parallel der optimierte MAPB-Postprozessor und das DD-Verfahren ihre Rechenschritte ausführen. Man beachte, obwohl die zweite Systemstufe auf die Berechnungen der erste Stufe angewiesen ist, entstehen dadurch keine nennenswerten Verzögerungen hinsichtlich der *online* Signalverarbeitung.

7.3. Experimentelle Untersuchungen

Während der nicht optimierte MAPB-Postprozessor aus Alg. 7.1 im Rahmen einer Machbarkeitsstudie in Unterabschnitt 7.1.2 in Kombination mit dem IMCRA-Verfahren experimentell untersucht wurde, soll der optimierte MAPB-Postprozessor aus Alg. 7.2 mit ver-

schiedenen Verfahren zur RLDS-Schätzung in unterschiedlichen Störumgebungen getestet werden. Dafür werden in dieser Arbeit zwei größere Experimentreihen mit der TIMIT-Datenbank und mit der CHiME-3-Datenbank durchgeführt[7]. In der ersten Experimentenreihe aus Unterabschnitt 7.3.1 soll die Leistungsfähigkeit des optimierten MAPB-Postprozessors beim Einsatz verschiedener konventioneller RLDS-Schätzer in der ersten Systemstufe für unterschiedliche Werte des eingangsseitigen $\mathrm{SNR}_{\mathrm{IN}}$ begutachtet werden. Hier werden sieben moderne RLDS-Schätzer eingesetzt, die in Abschnitt 6.3 beschrieben wurden. In den Untersuchungen auf den Daten der CHiME-3-Datenbank nehmen neben den erwähnten konventionellen Rauschschätzern auch die im Rahmen dieser Arbeit entwickelten RLDS-Schätzer aus Kap. 5 und Kap. 6 teil, sodass in der ersten Systemstufe insgesamt elf verschiedene Rauschschätzer getestet werden. Und da die konventionellen RLDS-Schätzer auf diesen Daten in Abschnitt 6.3 bereits optimiert wurden, werden sie hier sowohl mit den optimierten als auch mit den von den Entwicklern empfohlenen Parametern eingesetzt.

7.3.1. SNR-abhängige Ergebnisse auf TIMIT-Daten

Während konventionelle Rauschschätzer in der ersten Systemstufe aus Abb. 7.12 eingesetzt und entsprechend den Empfehlungen der Autoren parametrisiert werden, wird der MAPB-Postprozessor mit $\nu_0 = 40$, $\Delta\nu = 10$ und $\beta_{\max} = 0.01$ in der zweiten Systemstufe verwendet. Dabei soll die Leistungsfähigkeit des MAPB-Postprozessors hinsichtlich der Verfolgung von LDS des nichtstationären Rauschens und die daraus resultierenden Auswirkungen auf die Qualität der entstörten Sprachsignale untersucht werden. Wie in Unterabschnitt 7.1.2 soll auch hier analysiert werden, ob der optimierte MAPB-Postprozessor im Stande ist, die Schätzwerte der verwendeten konventionellen Rauschschätzer und somit auch die Qualität der entstörten Signale der ersten Systemstufe zu verbessern. Die hier vorgestellten Ergebnisse wurden zum Teil in [CHUTM13] präsentiert.

Die ungestörten Sprachsignale werden zunächst aus den Signalen der TIMIT-Datenbank zusammengestellt, die in Abschnitt 2.4 bereits beschrieben wurde. Bevor die kurzen Äußerungen der TIMIT-Datenbank zu einer langen Äußerung zusammengestellt werden, werden Sprachpausen am Anfang und am Ende jeder kurzen Äußerung entfernt. Die langen Äußerungen werden sowohl für die weiblichen Sprecher als auch für die männlichen Sprecher erstellt. Jede lange Äußerung dauert etwa 2 Minuten und beinhaltet Aufnahmen von 8 unterschiedlichen Sprechern. Am Anfang jeder langen Äußerung gibt es eine Zeitspanne von $0{,}1$ Sekunden ohne Sprachaktivität. Die ungestörten langen Sprachsignale werden mit drei unterschiedlichen nichtstationären Störsignalen additiv überlagert. Das erste Störsignal ist das hoch nichtstationäre Signal des Rauschtyps *babble* aus der NOISEX-92-Datenbank, die in Abschnitt 2.4 bereits beschrieben wurde. Das zweite ist das amplitudenmodulierte weiße normalverteilte Rauschen, das im Weiteren als ein sinusförmiges WNR (SWNR) bezeichnet wird. Seine Amplitude wird im Zeitbereich mit folgender Funktion moduliert

$$g(n) = 1 + \sin\left(2\pi \cdot \frac{f_{\mathrm{mod}}(n)}{F_{\mathrm{s}}} \cdot n\right) \tag{7.38}$$

mit einem Zeitindex n, mit einer Abtastrate F_{s} und mit einer zeitvarianten Modulationsfrequenz $f_{\mathrm{mod}}(n)$, die innerhalb von 30 Sekunden linear von $0\,\mathrm{Hz}$ auf $0.5\,\mathrm{Hz}$ ansteigt, wie

[7] An dieser Stelle darf nicht unerwähnt bleiben, dass der optimierte MAPB-Postprozessor auch für eine verbesserte modellbasierte ASR-Aufgabe eingesetzt werden kann, wie in [CPHU14] berichtet wird.

in [TTM+11]. Vom Höreindruck ähnelt dieser Rauschtyp einem Meeresrauschen mit einer leichten Brandung, bei der die Rate des Wellenganges mit der Zeit variiert. Das dritte Störsignal stammt aus der *SOUND-IDEAS*-Datenbank und beinhaltet Geräusche in einem Auto während der Beschleunigungs- und Abbremsvorgänge [Nim92]. Dieser Rauschtyp wird als *car* bezeichnet und soll ein nichtstationäres Rauschen milderen Grades repräsentieren. Alle verwendeten Signale weisen eine Abtastrate von $8\,\text{kHz}$ auf. Das globale SNR_{IN} wird zwischen $-5\,\text{dB}$ und $20\,\text{dB}$ in Schritten von $5\,\text{dB}$ variiert. Man beachte, obwohl die Leistungsfähigkeit der betrachteten RLDS-Schätzer für die empfohlenen Parameter bereits in Abschnitt 6.3 auf den CHiME-3-Daten analysiert wurde, soll die Untersuchung auf den TIMIT-Daten hier einen tieferen Einblick in das Verhalten der Verfahren bei verschiedenen Rauschtypen und bei unterschiedlichen SNR_{IN}-Werten bringen.

Als RLDS-Referenz $\lambda_{D,k,\ell}$ wird die zeitlich geglättete Version des momentanen wahren LDS des Störsignals $|D_{k,\ell}|^2$ verwendet. Im Unterschied zur Machbarkeitsstudie, die in Unterabschnitt 7.1.2 durchgeführt wurde, wird hier statt der rekursiven Glättung (7.22) mit einem Glättungsparameter $\alpha = 0{,}95$ eine verzögerungsfreie nichtkausale Glättung verwendet, die in *MATLAB* mit der Funktion *filtfilt*(1-α, [1 -α], x) berechnet wird. Die verzögerungsfreie Referenz verhindert eine Begünstigung der RLDS-Schätzer, welche eine ähnliche Verzögerung der Schätzwerte aufweisen wie eine verzögerungsbehaftete Referenz. Da die *filtfilt()* Funktion sowohl die Glättung in die Vorwärts- als auch in die Rückwärtsrichtung durchführt, wird α auf $0{,}9$ gesetzt. Die Glättung wird im STFT-Bereich in jedem Frequenzband über die ganze Signallänge durchgeführt. Bei der STFT wird die Quadratwurzel des *Hann*-Fensters und die FFT-Länge von 256 Abtastwerten verwendet. Der Rahmenvorschub wird entsprechend der Empfehlungen der Entwickler der Rauschschätzer gewählt. Während der Rahmenvorschub von $50\,\%$ der FFT-Länge für OSMS, MMSE-Hendriks und SNT-Rauschschätzer gewählt wird, beträgt er nur $25\,\%$ für MCRA-basierte Rauschschätzer. Die Verwendung unterschiedlicher Rahmenvorschübe resultiert in verschiedener Anzahl der berechneten RLDS-Schätzwerte. Um eine faire Bewertung der RLDS-Schätzer durchzuführen, bräuchte man in diesem Fall zwei unterschiedliche RLDS-Referenzen mit den gleichen Frequenzeigenschaften, die mit Hilfe von zwei unterschiedlichen jedoch aneinander angepassten Glättungskonstanten berechnet werden könnten. Im Rahmen dieser Bewertung wird allerdings eine einzige Referenz verwendet. Für einen fairen Vergleich der Schätzer werden in diesem Fall entbehrliche Schätzwerte der MCRA-basierten Verfahren einfach verworfen.

Da die verwendete Referenz $\lambda_{D,k,\ell}$ verzögerungsfrei ist, wird keine Verschiebung der Schätzwerte der Rauschleistungsdichte wie in den (7.25) mehr benötigt. Somit können der LEM-Schätzfehler und die LEV-Schätzfehlervarianz als Maße für die Leistungsfähigkeit der Rauschschätzer verwendet werden, wie sie in (2.43) und (2.45) definiert sind. Man beachte, dass die ersten drei Sekunden der Sprachsignale für die Initialisierung der Rauschschätzer verwendet werden und aus der Berechnung der beiden Bewertungsmaße ausgeschlossen sind. Außerdem fließen die fünf untersten und fünf obersten Frequenzbänder unterhalb der Nyquist-Frequenz in die Berechnung der Bewertungsmaße nicht ein, um den Einfluss des Hochpassfilters zum Entfernen des möglichen Gleichanteils im gestörten Sprachsignal und des Tiefpassfilters zum Vermeiden des *Aliasing*-Effektes zu reduzieren. Die Qualität der entstörten Signale wird mit den schmalbandigen MOS-LQO_{NB}-Werten bewertet. Alle Messgrößen werden über die Signale mit männlichen und weiblichen Sprechern gemittelt.

Bewertung der ersten Systemstufe: Die Leistungsfähigkeit der sieben in Abschnitt 3.1 beschriebenen Rauschschätzer, die in der ersten Systemstufe in Abb. 7.12 eingesetzt wer-

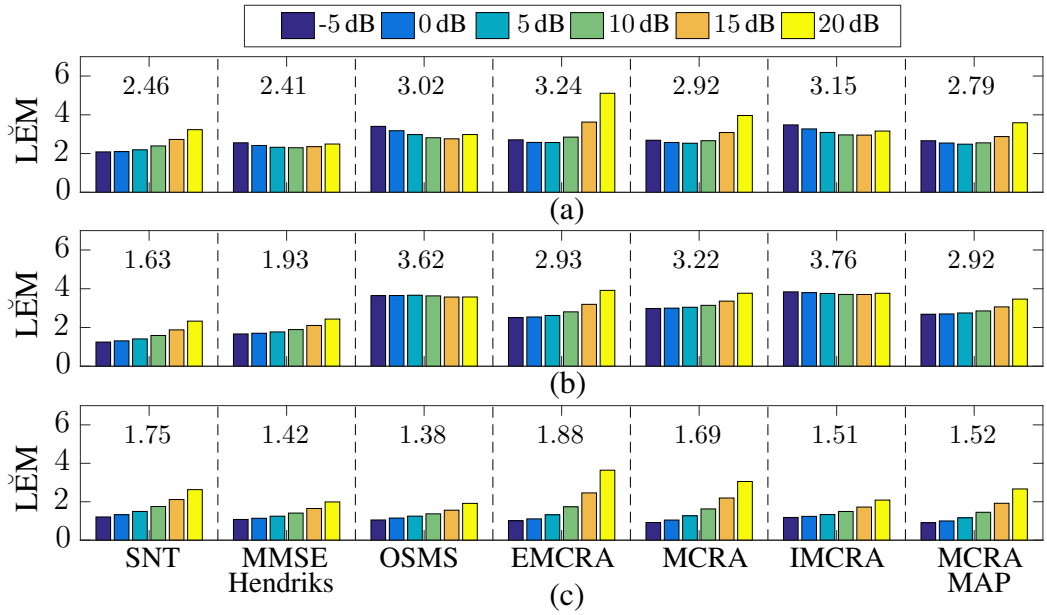


Abbildung 7.13.: LĔM-Werte der sieben verschiedenen RLDS-Schätzer, die in der ersten Stufe des Systems in Abb. 7.12 eingesetzt werden, für unterschiedliche $\mathrm{SNR}_{\mathrm{IN}}$-Werte und folgende nichtstationäre Rauschtypen: (a) *babble*, (b) SWNR, (c) *car*. Durchschnittliche LĔM-Werte gemittelt über alle simulierten $\mathrm{SNR}_{\mathrm{IN}}$-Werte sind separat angegeben.

den, ist in Abb. 7.13 für die drei betrachtete Störsignale und unterschiedliche Werte von $\mathrm{SNR}_{\mathrm{IN}}$ dargestellt. Die resultierenden LĔM-Werte werden aus den Schätzwerten $\breve{\lambda}_{D,k,\ell}$ berechnet. Um LEM-Werte verschiedener Störsignale untereinander besser vergleichen zu können, sind die y-Achsen gleich skaliert. Dabei fällt auf, dass die RLDS-Schätzung der mild nichtstationären Störsignale *car* allen Rauschschätzern generell etwas besser gelingt als bei den anderen Rauschtypen. Dieser Rauschtyp spiegelt am besten die Tatsache wider, dass der mittlere Schätzfehler aller Verfahren mit steigenden $\mathrm{SNR}_{\mathrm{IN}}$-Werten insgesamt ansteigt. Während alle Verfahren für kleine $\mathrm{SNR}_{\mathrm{IN}}$-Werte ähnliche Leistungsfähigkeit aufweisen, wächst der Schätzfehler mit steigenden $\mathrm{SNR}_{\mathrm{IN}}$-Werten bei solchen Verfahren wie MMSE-Hendriks, OSMS- und IMCRA schwächer als bei anderen Schätzern an. Also eignen sich diese Verfahren für die Schätzung der Rauschleistungsdichte des mild nichtstationären Störsignale am besten. Betrachtet man das OSMS- und das IMCRA-Verfahren in Gegenwart des SWNR-Rauschtyps, stellt man trotz ihrer Robustheit hinsichtlich verschiedener $\mathrm{SNR}_{\mathrm{IN}}$-Werte fest, dass sie im Mittel aufgrund der großen LEM-Werte unter allen Rauschschätzern am schlechtesten abschneiden. Im Unterschied dazu bleibt der MMSE-Schätzer von Hendriks neben dem SNT-Verfahren auch unter den besten Schätzverfahren. An dieser Stelle lohnt es sich, den Leser darauf aufmerksam zu machen, dass obwohl die MCRA-basierten Verfahren für kleine $\mathrm{SNR}_{\mathrm{IN}}$-Werte sehr gute Leistungsfähigkeit zeigen, ihnen ihre hohe Sensitivität hinsichtlich der $\mathrm{SNR}_{\mathrm{IN}}$-Werte zum Verhängnis wird. Ähnliche Konstellationen für die Rauschschätzer sind auch beim Störsignal *babble* zu beobachten, das bekanntlich hochgradig nichtstationär ist und aus diesem Grund für die Rauschschätzer noch schwerer als das nichtstationäre WNR-Rauschen zu verfolgen ist. Zusammenfassend folgt daraus, dass, wenn im Vordergrund die Schätzung der hoch nichtstationären Rauschtypen steht, der Rauschschät-

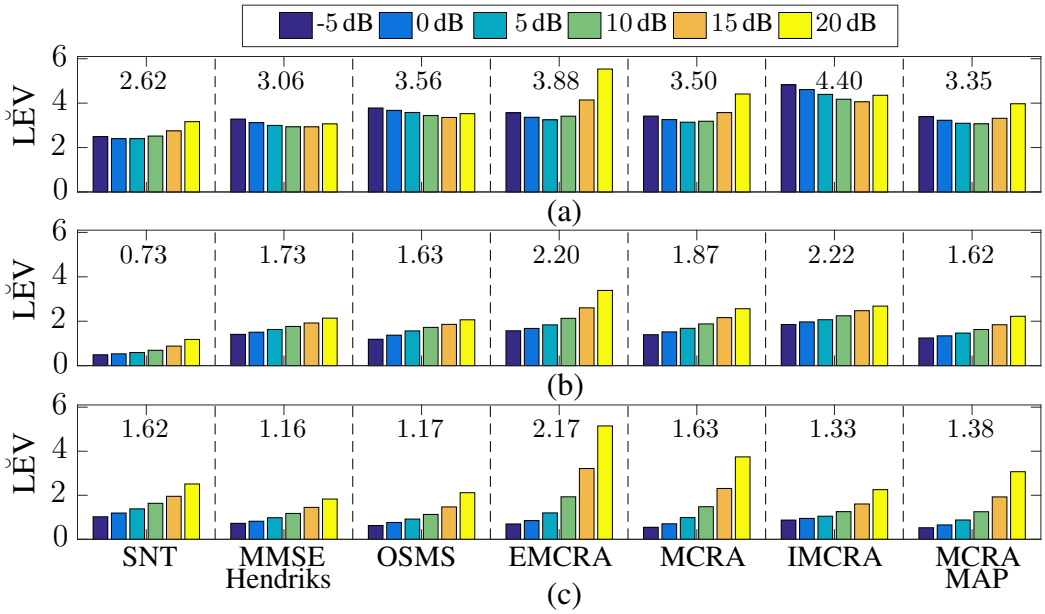


Abbildung 7.14.: LĔV-Werte der in Abschnitt 3.1 beschriebenen Rauschschätzer, die in der ersten Stufe des Systems in Abb. 7.12 eingesetzt werden, für unterschiedliche $\mathrm{SNR}_{\mathrm{IN}}$-Werte und Rauschtypen: (a) *babble*, (b) SWNR, (c) *car*. Durchschnittliche LĔV-Werte gemittelt über alle simulierten $\mathrm{SNR}_{\mathrm{IN}}$-Werte sind für jeden RLDS-Schätzer separat angegeben.

zer von MMSE-Hendriks und das SNT-Verfahren bevorzugt werden sollen[8]. Resultierende LĔV-Werte der Rauschschätzer sind in Abb. 7.14 dargestellt. Wie man sieht, sorgt hier der Rauschtyp *babble* für die größte mittlere Schätzfehlervarianz bei allen Rauschschätzern. Versucht man die besten Rauschschätzer unter Beachtung der beiden Messgrößen LEM und LEV zu bestimmen, sollen in Gegenwart des Rauschtyps *car* der Rauschschätzer MMSE-Hendriks und OSMS-Verfahren herausgestellt werden. Für das SWNR-Störsignal gewinnt eindeutig das SNT-Verfahren ein Duell mit dem MMSE-Hendriks Verfahren aufgrund der erstaunlich kleinen LEM- und LEV-Werte. Auch beim Störsignal *babble* behalten das SNT-Verfahren und der MMSE-Schätzer von Hendriks ihre Überlegenheit und schneiden hier in etwa gleich gut ab.

Eine gute RLDS-Schätzung dient allerdings keinem Selbstzweck sondern einem System zur Sprachsignalentstörung, dessen Ausgangssignale hinsichtlich ihrer Qualität gemessen in $\mathrm{MOS\text{-}LQO}_{\mathrm{NB}}$-Werten untersucht werden. Auswirkungen betrachteter RLDS-Schätzer auf die Sprachsignalqualität kann man anhand der Unterschiede der resultierenden $\mathrm{MOS\text{-}LQO}_{\mathrm{NB}}$-Werte gut sehen, die wie folgt definiert werden:

$$\Delta\breve{\mathrm{MOS\text{-}LQO}}_{\mathrm{NB}} = \breve{\mathrm{MOS\text{-}LQO}}_{\mathrm{NB}} - \overline{\mathrm{MOS\text{-}LQO}}_{\mathrm{NB}}, \tag{7.39}$$

wobei $\overline{\mathrm{MOS\text{-}LQO}}_{\mathrm{NB}}$ ein mittlerer schmalbandiger MOS-LQO-Wert für ein bestimmtes $\mathrm{SNR}_{\mathrm{IN}}$-

[8] Die Leistungsfähigkeit der betrachteten RLDS-Schätzer wird auch in [TTM+11] untersucht. Allerdings wird dort eine verzögerungsbehaftete Referenz $\lambda_{D,k,\ell}$ verwendet, die rekursiv mit $\alpha = 0{,}9$ berechnet wird. Vergleicht man die MSE-Werte aus Abb. 7.13 mit denen aus [TTM+11] für die Störsignale *babble* und SWNR, bleiben die Verhältnisse hinsichtlich der Leistungsfähigkeit der RLDS-Schätzer bis auf das IMCRA-Verfahren in etwa gleich. Gegenüberstellung der beiden Studien zeigt, dass das IMCRA-Verfahren durch die Verwendung der verzögerungsbehafteten Referenz für kleine $\mathrm{SNR}_{\mathrm{IN}}$-Werte stark begünstigt wird, obwohl seine Schätzwerte verglichen mit der verzögerungsfreien Referenz bei weitem nicht so gut sind.

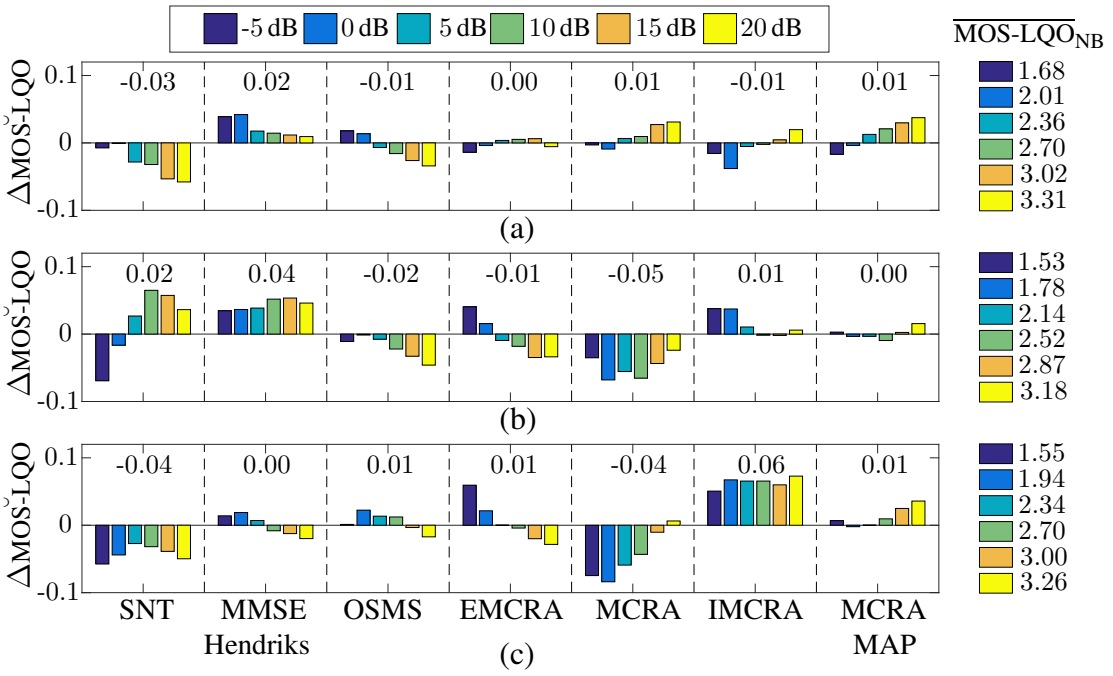


Abbildung 7.15.: $\Delta\text{MOS-LQO}_{\text{NB}}$- und $\overline{\text{MOS-LQO}}_{\text{NB}}$-Werte der betrachteten RLDS-Schätzer, die in der ersten Stufe des Systems in Abb. 7.12 eingesetzt werden, für unterschiedliche SNR_{IN}-Werte und Rauschtypen: (a) *babble*, (b) SWNR, (c) *car*.

Wert gemittelt über alle betrachteten RLDS-Schätzer ist. In anderen Worten geschieht die Berechnung von $\overline{\text{MOS-LQO}}_{\text{NB}}$ durch die Mittelung über die $\text{MOS-LQO}_{\text{NB}}$-Werte, die aus den Ausgangssignalen der ersten Systemstufe resultieren, in der unterschiedliche RLDS-Schätzer verwendet werden. Für unterschiedliche Störsignale und SNR_{IN}-Werte ergeben sich somit verschiedene $\overline{\text{MOS-LQO}}_{\text{NB}}$-Werte. Die mittleren $\overline{\text{MOS-LQO}}_{\text{NB}}$- und daraus resultierenden $\Delta\text{MOS-LQO}_{\text{NB}}$-Werte für unterschiedliche Störsignale, SNR_{IN}-Werte und betrachtete Rauschschätzer sind in Abb. 7.15 dargestellt.

Die $\overline{\text{MOS-LQO}}_{\text{NB}}$-Werte, die rechts von den Bildern in Abb. 7.15 platziert sind, zeigen, dass die Qualität der Ausgangssignale mit den steigenden SNR_{IN}-Werten bei allen Rauschtypen erwartungsgemäß wächst. Den $\Delta\text{MOS-LQO}_{\text{NB}}$-Werten ist zu entnehmen, dass die Verwendung von unterschiedlichen RLDS-Schätzern die Qualität der prozessierten Sprachsignale um etwa ± 0.05 Punkte des absoluten $\text{MOS-LQO}_{\text{NB}}$-Wertes beeinflusst, was einen Unterschied von etwa 0.1 ausmacht, der besonders für die stark gestörten Sprachsignale mit den kleinen SNR_{IN}-Werten signifikant ist. Beim Betrachten der experimentellen Ergebnisse der jeweiligen Rauschtypen fällt auf, dass der MMSE-Schätzer von Hendriks sich beim Rauschtyp *babble* besonders auszeichnet, wobei seine Stärken hier eindeutig bei den stark verrauschten Eingangssignalen liegen. Am schlechtesten schneidet hier der SNT-Schätzer ab, dessen Leistungsfähigkeit sich mit steigenden SNR_{IN}-Werten deutlich verschlechtert. Auch beim SWNR-Rauschen übertrifft der MMSE-Schätzer von Hendriks die anderen Rauschschätzer. Die hervorragende RLDS-Schätzung des SNT-Verfahrens trägt bei diesem Rauschtyp zu einer überdurchschnittlich guten Sprachsignalqualität bei. Die schlechteste Leistungsfähigkeit erreicht hier das MCRA-Verfahren. Nur beim Rauschtyp *car* kann sich der MMSE-Schätzer von Hendriks nicht auszeichnen und überlässt die Führung dem IMCRA-Schätzer,

der hier unabhängig von $\mathrm{SNR_{IN}}$-Werten überragend ist. Das SNT-Verfahren und der MCRA-Schätzer erreichen hier die niedrigste Qualität der prozessierten Sprachsignale. Obwohl das SNT-Verfahren bezüglich der RLDS-Schätzung unter den besten Rauschschätzern war, führt dieses Verfahren zu den Sprachsignalen mit mangelnder Qualität, was in der RLDS-Überschätzung begründet ist, die vom LEM-Bewertungsmaß nicht extra bestraft wird. Zusammenfassend, lässt sich festhalten, dass der Schätzer von Hendriks und das IMCRA-Verfahren gemittelt über alle betrachteten Rauschtypen und $\mathrm{SNR_{IN}}$-Werte die beste Arbeit hinsichtlich der Sprachsignalqualität leisten.

Bewertung des optimierten MAPB-Postprozessors als RLDS-Schätzer: Da allerdings die Verbesserung der RLDS-Schätzung und die daraus resultierende Verbesserung der Qualität der entstörten Signale der zweiten Stufe des Systems in Abb. 7.12 im Vergleich zur ersten Stufe in unseren Untersuchungen von Interesse sind, werden im Weiteren drei folgende differenzielle Bewertungsmaße betrachtet:

$$\Delta\mathrm{L\hat{E}M} = \mathrm{L\breve{E}M} - \mathrm{L\hat{E}M}, \qquad (7.40) \qquad\qquad \Delta\mathrm{L\hat{E}V} = \mathrm{L\breve{E}V} - \mathrm{L\hat{E}V}, \qquad (7.41)$$

$$\Delta\mathrm{MOS\hat{}\text{-}LQO_{NB}} = \mathrm{MOS\hat{}\text{-}LQO_{NB}} - \mathrm{MOS\breve{}\text{-}LQO_{NB}}, \qquad (7.42)$$

wobei $\mathrm{L\breve{E}M}$, $\mathrm{L\breve{E}V}$ und $\mathrm{MOS\breve{}\text{-}LQO_{NB}}$ aus den Schätzwerten $\breve{\lambda}_{D,k,\ell}$ und $\breve{s}(t)$ der ersten Systemstufe und $\mathrm{L\hat{E}M}$, $\mathrm{L\hat{E}V}$ und $\mathrm{MOS\hat{}\text{-}LQO_{NB}}$ aus den Schätzwerten $\hat{\lambda}_{D,k,\ell}$ und $\hat{s}(t)$ der zweiten Stufe berechnet werden, in welcher der MAPB-Postprozessor als ein RLDS-Schätzer eingesetzt wird. Während $\Delta\mathrm{L\hat{E}M}$ und $\Delta\mathrm{L\hat{E}V}$ die Reduktion des mittleren Schätzfehlers angeben, die vom Postprozessor erreicht wird, drückt $\Delta\mathrm{MOS\hat{}\text{-}LQO_{NB}}$ die hierdurch entstandene Qualitätsverbesserung der entstörten Signale aus. Alle differenziellen Messgrößen aus (7.40)-(7.42) sind so definiert, dass größere Werte bessere Leistungsfähigkeit zeigen.

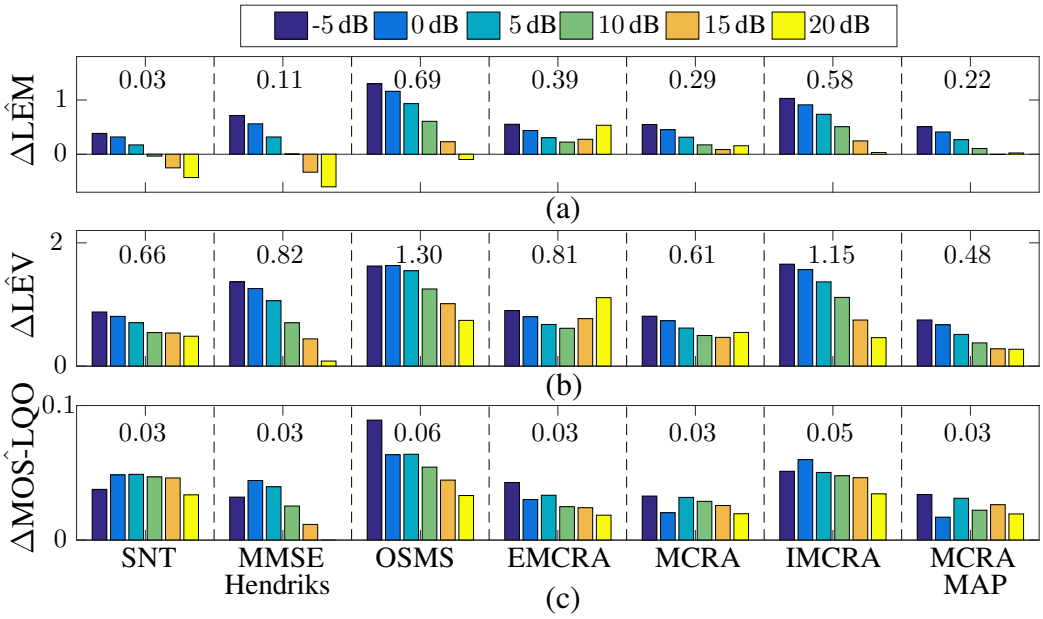


Abbildung 7.16.: Einfluss des optimierten MAPB-Postprozessors eingesetzt im System in Abb. 7.12 auf Rauschschätzung und Sprachsignalqualität für den Rauschtyp *babble* bei unterschiedlichen $\mathrm{SNR_{IN}}$-Werten: (a) $\Delta\mathrm{L\hat{E}M}$, (b) $\Delta\mathrm{L\hat{E}V}$, (c) $\Delta\mathrm{MOS\hat{}\text{-}LQO_{NB}}$. Durchschnittliche Bewertungsmaße gemittelt über alle simulierten $\mathrm{SNR_{IN}}$-Werte sind separat angegeben.

In Abb. 7.16 ist der Einfluss des MAPB-Postprozessors auf Rauschschätzung und Signalentstörung für den Rauschtyp *babble* bei unterschiedlichen SNR_{IN}-Werten dargestellt. Betrachtet man die durchschnittlichen Werte von $\Delta\hat{\text{LEM}}$, $\Delta\hat{\text{LEV}}$ und $\Delta\hat{\text{MOS-LQO}}$, die für jeden RLDS-Schätzer aus der ersten Stufe separat angegeben sind, fällt auf, dass fast alle Deltagrößen positiv sind. Dies ist ein Indiz dafür, dass der MAPB-Postprozessor nicht nur die RLDS-Schätzung aller Verfahren verbessert, sondern überall auch zur Verbesserung der Sprachsignalqualität beiträgt. Dabei ist zu beachten, dass der Rauschtyp *babble* sich aufgrund seiner hohen Nichtstationarität als eine der schwierigsten Störumgebungen für eine RLDS-Schätzung darstellt, siehe Abb. 7.13 und Abb. 7.14. Gemittelt über alle betrachteten RLDS-Schätzer der ersten Systemstufe werden das LEM-Maß von 2.86 um 0.33 Punkte also etwa $11.6\,\%$ und das LEV-Maß von 3.48 um 0.83 Punkte also etwa $23.9\,\%$ reduziert, siehe Abb. 7.16 (a) und (b). Dabei verbessert der MAPB-Postprozessor die RLDS-Schätzung besonders bei den sehr stark verrauschten Signalen mit kleinen SNR_{IN}-Werten. Genauere RLDS-Schätzung führt auch zur Verbesserung der Sprachsignalqualität, die umso höher ausfällt, je stärker die Sprachsignale verrauscht sind, siehe Abb. 7.16 (c). So werden die $\text{MOS-LQO}_{\text{NB}}$-Werte der ersten Systemstufe für $\text{SNR}_{\text{IN}} = \text{-}5\,\text{dB}$ von 1.677 um 0.046 Punkte erhöht, was prozentuell etwa $6.7\,\%$ beträgt, wenn man berücksichtigt, dass das $\text{MOS-LQO}_{\text{NB}}$-Maß den minimalen Wert von 1 hat. Differenziert man zwischen den einzelnen RLDS-Schätzern, fällt es auf, dass das OSMS- und das IMCRA-Verfahren eingesetzt in der ersten Systemstufe vom MAPB-Postprozessor am meisten profitieren und zwar in allen betrachteten Bewertungsmaßen. Und obwohl das LEM-Maß des MMSE-Schätzers von Hendriks und des SNT-Verfahrens im Mittel nur minimal reduziert wird, sorgt eine signifikante Verringerung des LEV-Maßes hier für eine eindeutige Verbesserung der Sprachsignalqualität. Betrachtet man Ergebnisse des IMCRA-Verfahrens detaillierter und vergleicht sie mit Resultaten in Abb. 7.4 und in Abb. 7.5, stellt man fest, dass die Optimierung des MAPB-Postprozessors, die in Unterabschnitt 7.2.2 vorgestellt wurde, seine Eigenschaften für gestörte Sprachsignale mit höheren SNR_{IN}-Werten verbesserte.

Um den durchschnittlichen Einfluss des optimierten MAPB-Schätzers auf Rauschschätzung und Sprachsignalqualität für verschiedene Rauschtypen zu verdeutlichen, sind die berechneten Bewertungsmaße gemittelt über alle SNR_{IN}-Werte und alle verwendeten RLDS-Schätzer in Tab. 7.1 zusammengefasst. Dabei resultieren die eingetragenen Werte $\breve{\text{LEM}}$, $\breve{\text{LEV}}$ und $\breve{\text{MOS-LQO}}$ aus den Ergebnissen, die in Abb. 7.13, in Abb. 7.14 und in Abb. 7.15 vorgestellt wurden. Obwohl die betrachteten RLDS-Schätzer für die Rauschtypen *babble* und SWNR die gleichen $\breve{\text{LEM}}$-Werte erreichen, fallen die $\breve{\text{LEV}}$ Maße hier zugunsten des SWNR-Rauschens aus, das eine kleinere Schätzfehlervarianz hervorruft. Ungeachtet dessen, dass die RLDS-Schätzung hier besser als beim *babble* Rauschen gelingt, ist die spektrale Sprachsi-

-	$\breve{\text{LEM}}$	$\Delta\hat{\text{LEM}}$	$\breve{\text{LEV}}$	$\Delta\hat{\text{LEV}}$	$\breve{\text{MOS-LQO}}_{\text{NB}}$	$\Delta\hat{\text{MOS-LQO}}_{\text{NB}}$
babble	2.86	0.33	3.48	0.83	2.52	0.04
SWNR	2.86	0.36	1.71	0.35	2.34	0.04
car	1.59	-0.03	1.50	0.18	2.47	0

Tabelle 7.1.: Durchschnittlicher Einfluss des optimierten MAPB-Postprozessors eingesetzt in Abb. 7.12 auf Rauschschätzung und Sprachsignalentstörung für verschiedene Rauschtypen.

gnalentstörung dieses Rauschtyps etwas erschwert, worauf ein Vergleich der MOS$\breve{}$-LQO-Werte hinweist. Jedoch verbessert der MAPB-Postprozessor auch bei diesem Rauschtyp die Genauigkeit der RLDS-Schätzung deutlich, die sich auch auf die Qualität der Ausgangssignale positiv auswirkt. Die resultierenden ΔMOS$\hat{}$-LQO-Werte fallen hier dadurch relativ bescheiden aus, dass der MAPB-Postprozessor die Signalqualität der ersten Systemstufe mit dem SNT-Verfahren und dem MMSE-Schätzer von Hendriks, die beim SWNR-Rauschen bereits eine sehr gute Genauigkeit aufweisen, nicht verbessern konnte. Im Gegensatz zu diesen beiden Verfahren profitierten der OSMS-Schätzer und alleMCRA-basierten Verfahren stark von der Verwendung der zweiten Systemstufe. Wie die resultierenden LE$\breve{}$M- und LE$\breve{}$V-Werte zeigen, meistern die RLDS-Schätzer in Gegenwart des Rauschtyps *car* ihre Aufgabe bereits gut genug, sodass der MAPB-Postprozessor nur die Schätzfehlervarianz leicht reduzieren kann. Wichtig ist festzuhalten, dass die zweite Systemstufe die Sprachsignalqualität hier nicht verschlechtert.

Zusammenfassend lässt sich festhalten, dass der optimierte MAPB-Postprozessor zur Verbesserung der Leistungsfähigkeit der spektralen Sprachsignalentstörung in Gegenwart von stark nichtstationären Rauschtypen beiträgt, die in der Praxis häufig auftreten. Dabei wird die gute Leistungsfähigkeit der konventionellen RLDS-Schätzer bei stationären Rauschtypen aufrechterhalten. Außerdem verbessert die in Unterabschnitt 7.2.2 durchgeführte Optimierung die Eigenschaften des MAPB-Postprozessors bei leicht verrauschten Sprachsignalen mit kleinen $\mathrm{SNR_{IN}}$-Werten. Die zweite Systemstufe, in der der MAPB-Postprozessor eingesetzt wird, kann effizient realisiert werden.

7.3.2. Leistungsfähigkeit auf CHiME-3-Daten

Im Unterschied zu den Untersuchungen in Unterabschnitt 7.1.2 und Unterabschnitt 7.3.1 wird im Rahmen dieser Auswertung jede der beiden Systemstufen so wie das in Abschnitt 6.3 beschriebene System zur spektralen Entstörung aufgebaut, also ohne den Baustein 4 aus Abb. 2.2. Somit wird gezeigt, dass der MAPB-Postprozessor auch in den Systemen zur spektralen Sprachsignalentstörung eingesetzt werden kann, die anders als in Unterabschnitt 7.1.2 oder Unterabschnitt 7.3.1 konfiguriert werden. Da die Parameter der in der ersten Systemstufe eingesetzten RLDS-Schätzer in Abschnitt 6.3 auf den $25\,\%$ der CHiME-3-Daten optimiert wurden, sollen auch Parameter des optimierten MAPB-Postprozessors auf denselben Daten optimal gewählt werden. Und da der MAPB-Postprozessor sich laut den früheren Untersuchungen besonders effizient hinsichtlich der Verbesserung der Qualität der prozessierten Sprachsignale erwies, sollen seine Parameter so gesetzt werden, dass die resultierenden entstörten Sprachsignale auf den Optimierungsdaten die höchsten $\mathrm{MOS\text{-}LQO_{WB}}$-Werte erreichen. Da der Regelbereich des Freiheitsgrades $\Delta\nu$ aus Alg. 7.2 laut den Untersuchungen eine untergeordnete Rolle hinsichtlich der Leistungsfähigkeit des MAPB-Postprozessors spielt, wird er über den Zusammenhang $\Delta\nu = \frac{\nu_0}{4}$ in Abhängigkeit vom Freiheitsgrad gesetzt, sodass nur der Freiheitsgrad ν_0 und der Biaskompensationsfaktor β_{max} optimiert werden müssen. Und da die unterschiedlich parametrisierten RLDS-Schätzer der ersten Systemstufe durchaus verschiedene Eigenschaften aufweisen, werden die Parameter ν_0 und β_{max} des MAPB-Postprozessors für jeden der verwendeten Schätzer separat optimiert. Dabei werden die RLDS-Schätzer der ersten Stufe wie in Abschnitt 6.3 sowohl mit den von den Entwicklern empfohlenen (EMP) als auch mit den optimierten (OPT) Parametern parametrisiert, was zu unterschiedlichen Parametern ν_0 und β_{max} führt.

RLDS-Schätzer der ersten Stufe	EMP		OPT		RLDS-Schätzer der ersten Stufe	EMP		OPT	
	ν_0	β_{max}	ν_0	β_{max}		ν_0	β_{max}	ν_0	β_{max}
OSMS	25	0.01	30	0.075	BSMS	20	0	30	0.04
VAD-RA	15	0	40	0.125	SPP-FP	15	0.01	30	0.05
MCRA	30	0.01	40	0.125	EMCRA	15	0	40	0.125
MCRA-MAP	10	0	30	0.125	IMCRA	30	0.03	35	0.05
MMSE-VAD	20	0	30	0.05	MMSE-BM	25	0.05	30	0.05

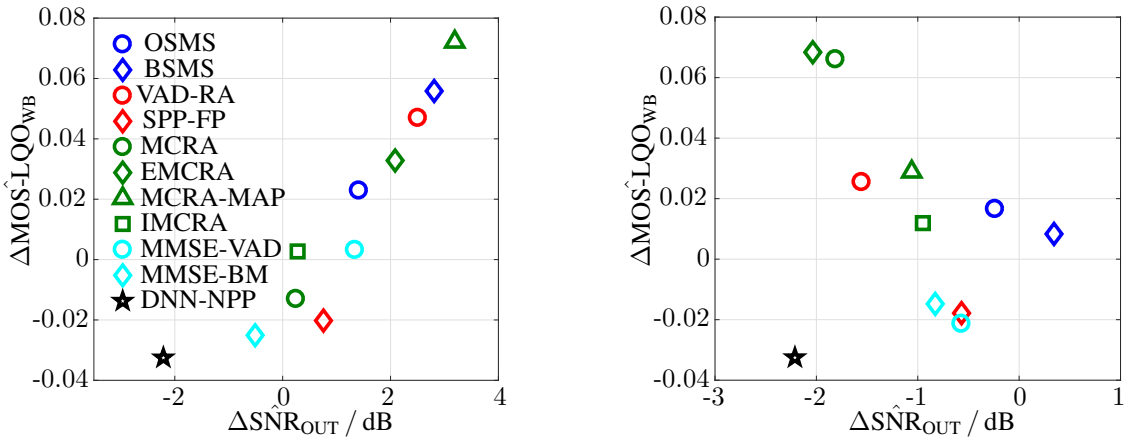
Tabelle 7.2.: Parameter des MAPB-Postprozessors ν_0 und β_{max} optimiert auf CHiME-3-Optimierungsdaten für unterschiedliche RLDS-Schätzer der ersten Systemstufe, die mit empfohlenen (EMP) und optimierten (OPT) Parametern parametrisiert werden.

In Tab. 7.2 sind die optimalen MAPB-Parameter angegeben, die sich auf den Optimierungsdaten ergeben. Beim genaueren Betrachten der resultierenden Freiheitsgrade ν_0 fällt auf, dass die RLDS-Schätzer der ersten Systemstufe mit empfohlenen Parametern die Freiheitsgrade im Bereich $[10; 30]$ und die mit optimierten Parametern im Bereich $[30; 40]$ benötigen. Daraus folgt, dass eine kleinere Bandbreite des MAPB-Postprozessors bei den optimierten Parametern erforderlich ist als dies bei den empfohlenen Parametern der Fall ist. Um die beste Sprachsignalqualität zu erreichen, soll der MAPB-Postprozessor demnach eine stärkere Glättung bei den optimierten Parametern realisieren als bei den empfohlenen. Erwähnenswert ist die Tatsache, dass der in Unterabschnitt 7.2.2 empfohlene Freiheitsgrad von $\nu_0 = 40$, der sich aus einer nummerischen Qualitätsanalyse des MAPB-Postprozessors ergab, in etwa den Freiheitsgraden aus Tab. 7.2 für optimierte Parameter entspricht.

Die Begutachtung der resultierenden Biaskompensationsfaktoren β_{max} offenbart eine weitere bemerkenswerte Gesetzesmäßigkeit. Die RLDS-Schätzer der ersten Stufe mit den empfohlenen Parametern benötigen relativ kleine Biaskompensationsfaktoren im Wertebereich $[0; 0.05]$. Gemittelt über alle betrachteten RLDS-Schätzer bekommt man hier einen Biaskompensationsfaktor von etwa 0.011, der ziemlich genau mit dem in Unterabschnitt 7.2.2 empfohlenen Wert von β_{max} übereinstimmt, siehe Abb. 7.10. Werden bei den RLDS-Schätzern der ersten Systemstufe die optimierten Parameter verwendet, muss der Biaskompensationsfaktor des MAPB-Postprozessors in der Regel erhöht werden, um die beste Sprachsignalqualität zu erreichen. Hier liegt β_{max} im Wertebereich $[0.04; 0.125]$ mit dem Durchschnittswert von 0.08, wenn man über alle RLDS-Schätzer in Tab. 7.2 mittelt. Im Wesentlichen sorgen diese relativ großen Biaskompensationsfaktoren dafür, dass in der Sprachsignalaktivität keine RLDS-Überschätzung stattfindet, die bei der spektralen Sprachsignalentstörung zu einer fälschlicherweisen Unterdrückung von Signalanteilen des ungestörten Sprachsignals führt und somit sich negativ auf die Sprachsignalqualität auswirkt. Ein weiteres erwähnenswertes Ergebnis der Optimierung der MAPB-Parameter betrifft die Verwendung des DNN-basierten RLDS-Schätzers aus Kap. 6 in der ersten Systemstufe. In diesem Fall bekommt man $\nu_0 = 45$ und $\beta_{\text{max}} = 0.125$ als optimale MAPB-Parameter, die in Tab. 7.2 nicht angegeben sind.

Die Leistungsfähigkeit des MAPB-Postprozessors auf den CHiME-3-Evaluierungsdaten, die in Abschnitt 6.4 beschrieben werden, ist in Abb. 7.17 dargestellt. Hier ist die Verbesserung der Sprachsignalqualität gemessen in $\Delta\text{MOS}\hat{}\text{-LQO}_{\text{WB}}$-Werten, die ähnlich wie in (7.42) definiert werden, über die Erhöhung der Störsignaldämpfung $\Delta\text{S}\hat{\text{N}}\text{R}_{\text{OUT}}$ dargestellt mit

$$\Delta\text{S}\hat{\text{N}}\text{R}_{\text{OUT}} = \text{S}\hat{\text{N}}\text{R}_{\text{OUT}} - \text{S}\breve{\text{N}}\text{R}_{\text{OUT}}, \tag{7.43}$$



(a) Empfohlene Parameter

(b) Optimale Parameter

Abbildung 7.17.: Leistungsfähigkeit des MAPB-Postprozessors auf CHiME-3-Evaluierungsdaten beim Einsatz verschiedener RLDS-Schätzer in der ersten Systemstufe: (a) von den Entwicklern empfohlene Parameter, (b) optimierte Parameter. Der MAPB-Postprozessor wird entsprechend Tab. 7.2 parametrisiert.

wobei $\widehat{\text{SNR}}_{\text{OUT}}$ und $\breve{\text{SNR}}_{\text{OUT}}$ die Störsignaldämpfung jeweils der ersten und der zweiten Systemstufen sind. $\Delta\widehat{\text{MOS-LQO}}_{\text{WB}}$ und $\Delta\widehat{\text{SNR}}_{\text{OUT}}$ geben also die Verbesserung der Leistungsfähigkeit der zweiten Systemstufe gegenüber der ersten an.

Parametrisiert man die betrachteten RLDS-Schätzer mit den von jeweiligen Autoren empfohlenen Parametern, kann vom MAPB-Postprozessor für MCRA-MAP, BSMS, VAD-RA, EMCRA und OSMS-Rauschschätzer sowohl eine Verbesserung der Sprachsignalqualität als auch eine Erhöhung der Störsignaldämpfung erreicht werden, wie man in Abb. 7.17 (a) sieht. Dabei werden ähnliche oder sogar etwas größere Verbesserungen der Sprachsignalqualität erreicht als in Tab. 7.1 für *babble* und SWNR-Rauschtypen. Hierzu soll man hinzufügen, dass die erste Systemstufe mit diesen RLDS-Schätzern bei den Experimenten auf den CHiME-3-Evaluierungsdaten viel niedrigere MOS-LQO-Werte von etwa 1.37 liefert, wie man in Abb. 6.4 sieht, als auf den TIMIT-Daten. Man beachtet man außerdem, dass die erste Systemstufe gemittelt über diese Schätzer ein durchschnittliches ausgangsseitiges SNR_{OUT} von 8 dB liefert, gewinnt die Vergrößerung der Störsignaldämpfung um durchschnittlich etwa 2.4 dB an Wichtigkeit. Bei den MCRA-, IMCRA-, MMSE-VAD- und SPP-FP-Verfahren sind zwar $\Delta\widehat{\text{SNR}}_{\text{OUT}}$-Werte positiv, jedoch bis auf die letzten beiden Verfahren nicht besonders auffallend. Während beim MMSE-VAD-Schätzer die Vergrößerung des SNR_{OUT}-Wertes ohne Einbuße in der Sprachsignalqualität erreicht wird, führt die bessere Störsignaldämpfung beim SPP-FP-Verfahren zu den leichten Verlusten in der Sprachsignalqualität. Die Leistungsfähigkeit der ersten Systemstufe mit den beiden verbleibenden Schätzern MMSE-BM und DNN-NPP kann vom MAPB-Postprozessor leider nicht verbessert werden, denn sie liefern auf den CHiME-3-Evaluierungsdaten bereits eine gute RLDS-Schätzung und Signalentstörung, siehe Abb. 6.4.

Optimiert man die Parameter der RLDS-Schätzer in der ersten Systemstufe ergibt sich eine etwas andere Konstellation der resultierenden Bewertungsmaße, die in Abb. 7.17 (b)

dargestellt sind. Bei den meisten RLDS-Schätzern führt der MAPB-Postprozessor auch hier wieder zur Verbesserung der Sprachsignalqualität der prozessierten Signale. Bemerkenswert sind die $\Delta\text{MOS}\hat{}\text{-LQO}_{\text{WB}}$-Werte, die beim Einsatz von MCRA- und EMCRA-Verfahren erreicht werden. Allerdings geht diese Verbesserung auf Kosten der Reduzierung der Störsignaldämpfung von etwa -2 dB. Dabei fällt der Zusammenhang auf, dass je kleiner die Verbesserung der Sprachsignalqualität ist desto kleiner sind auch die Verluste in der Störsignaldämpfung. Eine einzige positive Ausnahme stellt hier der in Kap. 5 entwickelte BSMS Schätzer dar, der in Kombination mit dem MAPB-Postprozessor sowohl eine leichte Verbesserung in der Sprachsignalqualität als auch eine um 0.34 dB bessere Störsignaldämpfung erreicht. Im Gegensatz dazu verschlechtert der MAPB-Postprozessor die Leistungsfähigkeit der beiden MMSE-basierten Verfahren und des SPP-FP-Schätzers, die sich unter den konventionellen RLDS-Schätzern auf den CHiME-3-Evaluierungsdaten als die drei besten RLDS-Schätzer darstellen, siehe Abb. 6.4. Sonst leidet unter der Verwendung des MAPB-Postprozessors der DNN-NPP-Schätzer am meisten. Aus der Auswertung in Abb. 7.17 (b) lässt sich schlussfolgern, dass der MAPB-Postprozessor hier leider keinen großen Spielraum für die Verbesserung der gesamten Leistungsfähigkeit eines Systems zur spektralen Entstörung hat, denn der DNN-NPP-Schätzer bereits sehr gute Schätzwerte liefert.

7.4. Zusammenfassung

In diesem Kapitel wurde ein Bayessches Verfahren zur RLDS-Schätzung entwickelt, das in einem System zur spektralen Sprachsignalentstörung als Postprozessor agiert. Der vorgeschlagene MAP-basierte Postprozessor wurde in zwei Versionen vorgestellt, und zwar nichtoptimiert und optimiert. Ein Einsatz des nichtoptimierten MAPB-Postprozessors in der spektralen Sprachsignalentstörung zeigte, dass das vorgeschlagene Verfahren zur Reduktion der Schätzfehlervarianz der RLDS-Schätzung führt und somit eine Verbesserung der Qualität der entstörten Signale bewirkt. Eine numerische Qualitätsanalyse des nichtoptimierten MAPB-Postprozessors deckte jedoch einige Unzulänglichkeiten des Verfahrens auf, welche durch eine Biaskorrektur und eine Bandbreitenanpassung weitgehend beseitigt wurden. Dadurch entstand die optimierte Version des MAPB-Schätzers, dessen Leistungsfähigkeit in den umfangreichen Experimenten mit insgesamt zwölf verschiedenen RLDS-Schätzern in der ersten Systemstufe untersucht wurde.

Die Untersuchungen auf den Daten der TIMIT-Datenbank zeigten, dass der MAPB-Postprozessor die RLDS-Schätzung der ersten Systemstufe sowohl hinsichtlich des mittleren logarithmischen Schätzfehlers als auch bezüglich der mittleren logarithmischen Schätzfehlervarianz präzisieren kann, insbesondere bei solchen nichtstationären Störungen wie *babble noise*. Genauere RLDS-Schätzung sorgt dabei für eine leichte Verbesserung der Qualität der prozessierten Sprachsignale, die um so größer ausfällt, je stärker Signale verrauscht sind. Auch auf den CHiME-3-Daten wurde eine Verbesserung der Sprachsignalqualität beobachtet, die allerdings in manchen Fällen mit leichten Verlusten in der Störsignaldämpfung einherging. Die einzige Ausnahme hier stellten der DNN-basierte RLDS-Schätzer und das MMSE-BM-Verfahren dar, welche bereits eine gute Leistungsfähigkeit lieferten.

C. GENERALISIERTE MODELL-BASIERTE ENTSTÖRUNG

8. MAP-Schätzer generalisierter log-spektraler Amplituden

Wie in Abschnitt 2.2 beschrieben, werden Eigenschaften der entstörten Signale sehr stark von der Wahl der spektralen Filterfunktion $G(k, \ell)$ bestimmt. In Abschnitt 3.2 wurde eine Reihe von generalisierten modellbasierten spektralen Funktionen diskutiert, die neben den weit verbreiteten konventionellen Filterfunktionen existieren. Bei der Herleitung solcher Funktionen werden häufig entweder generalisierte spektrale Amplituden (GSA) wie des ungestörten Sprachsignals $|S(k, \ell)|^\beta$ mit einem Kompressionsfaktor $\beta \in \mathbb{R}_{>0}$ betrachtet oder generalisierte Verteilungsdichtefunktionen wie eine Gamma-, Chi- oder generalisierte Gamma-Verteilung eingesetzt, wie man Tab. 3.2 entnehmen kann. Allerdings gelang es den generalisierten Filterfunktionen bis jetzt noch nicht, den etablierten konventionellen Filterfunktionen, wie dem Wiener-Filter, dem MMSE-LSA-Schätzer und der OMLSA-Filterfunktion, eine ernsthafte Konkurrenz zu machen. Einige Gründe dafür sind die Notwendigkeit von spezifischen mathematischen Funktionen, mangelhafte Leistungsfähigkeit in der Sprachsignalqualität oder der Bedarf von zusätzlichen Systemkomponenten. Wie in Abschnitt 4.2 motiviert, wird im Rahmen dieses Kapitels eine neuartige generalisierte spektrale Filterfunktion vorgestellt, die zum Einen einfach und leistungsfähig ist und zum Anderen keine zusätzlichen Systemkomponenten benötigt. Um einen besseren Kompromiss zwischen der Sprachsignalqualität und der Störsignaldämpfung zu erreichen, kombiniert die neue Filterfunktion sowohl einen Schätzer im Bereich der log-spektralen Amplituden, welcher eine gute Qualität der prozessierten Sprachsignale mit sich bringt, als auch einen generalisierten Schätzer wie in [STCT98], welcher für eine gute Störsignaldämpfung sorgt. Diese Kombination wird durch eine neue Nichtlinearität realisiert, welche in den Bereich der logarithmischen GSA führt.

Der vorgeschlagene LGSA-Schätzer wird in Abschnitt 8.1 in fünf Schritten analytisch hergeleitet. In Unterabschnitt 8.1.1 werden dabei die ersten beiden Schritte beschrieben: statistische Modellierung im GSA-Bereich und Approximation dieser Modelle. Im Rahmen des dritten Schrittes in Unterabschnitt 8.1.2 wird die bedingte Verteilungsdichtefunktion im GSA-Bereich hergeleitet, die zum bekannten MMSE-PGSS-Schätzer aus [STCT98] führt. In Unterabschnitt 8.1.3 werden anschließend die letzten beiden Herleitungsschritte beschrieben: Trunkierung der bedingten VDF und Transformation in den LGSA-Bereich. Um mögliche Diskrepanzen zwischen den verwendeten Modellen und den Daten zu eliminieren, wird in Abschnitt 8.2 eine Anpassungsstrategie der Filterfunktion an die Daten vorgeschlagen, welche in einer datengetriebenen Parametrisierung der LGSA-Filterfunktion mündet. In Abschnitt 8.3 wird die Leistungsfähigkeit der vorgeschlagenen Filterfunktion im Vergleich zu den Filterfunktionen des weit verbreiteten MMSE-LSA-Schätzers aus [EM85] und des generalisierten MMSE-PGSS-Schätzers aus [STCT98] auf den CHiME-3-Daten ausgewertet.

Zum Schluss dieses Kapitels wird in Abschnitt 8.4 eine kurze Zusammenfassung gegeben. Man beachte, dass die in diesem Kapitel zu entwickelnde LGSA-Filterfunktion vom Autor dieser Arbeit zum ersten Mal in [CHU17] vorgestellt wurde.

8.1. Schätzer der logarithmischen generalisierten spektralen Amplituden

Bevor die vorgeschlagene LGSA-Filterfunktion hergeleitet wird, soll zunächst ihre Stellung im Bezug auf einige der generalisierten Filterfunktionen geklärt werden, die in Abschnitt 3.2 bereits beschrieben wurden. Dafür werden in Tab. 8.1 neben der LGSA-Filterfunktion vier weitere generalisierte Filterfunktionen aus Tab. 3.2 aufgelistet, die in einer Relation zur vorgeschlagenen Filterfunktion stehen. So ergibt sich der MMSE-Schätzer aus [STCT98], der so wie der LGSA-Schätzer bei seiner Herleitung von den generalisierten spektralen Amplituden Gebrauch macht, als ein Zwischenergebnis bei Herleitung der LGSA-Filterfunktion, wie im Weiteren gezeigt wird. Der Schätzer aus [DTI05] ist genauso wie der vorgeschlagene Schätzer ein MAP-basierter Schätzer und verwendet außerdem eine generalisierte Verteilungsdichtefunktion. Wie auch der LGSA-Schätzer entsteht die Filterfunktion aus [BKM08] dadurch, dass sowohl die generalisierten spektralen Amplituden als auch generalisierten Verteilungen verwendet werden. Der Schätzer aus [BA11] wird genauso wie der zu entwickelnde Schätzer im Bereich der log-spektralen Amplituden aufgestellt. Doch trotz einiger Gemeinsamkeiten unterschiedet sich der vorgeschlagene LGSA-Schätzer von allen Filterfunktionen aus Tab. 8.1. Seine besonderen Merkmale sind statistische Modellierung mit Hilfe einer Weibull-Verteilung und die Verwendung einer neuartigen Nichtlinearität.

Einführung einer neuartigen Nichtlinearität: Wie in [CG08] eingeführt, werden MMSE-basierte spektrale Filterfunktionen ausgehend von folgender Gleichung hergeleitet

$$f(\hat{S}(k,\ell)) = \mathbb{E}[\, f(S(k,\ell))\,|Y(k,\ell)], \tag{8.1}$$

	Quelle	Gen. Amplitude	Gen. Verteilung	Gen. Vorschrift	MMSE	MAP	Rayleigh	Gamma	Chi	Gen. Gamma	Weibull	Log-Schätzer
1.	[STCT98]	✓			✓		✓					
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
3.	[DTI05]		✓			✓				✓		
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
8.	[BKM08]	✓	✓		✓				✓			
9.	[BA11]		✓		✓			✓	✓			✓
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
12.	LGSA	✓	✓			✓					✓	✓

Tabelle 8.1.: Eigenschaften der vorgeschlagenen LGSA-Filterfunktion.

wobei die Funktion $f(x)$ sich als die verwendete Nichtlinearität darstellt, wenn man den Sonderfall $f(x) = x$ außer Acht lässt. Während die generalisierten Filterfunktionen häufig von der Nichtlinearität $f(x) = |x|^\beta$ Gebrauch machen, verwenden die konventionellen spektralen Filterfunktionen solche Nichtlinearitäten wie $f(x) = |x|$ oder $f(x) = \ln|x|$. So benutzt der MMSE-LSA-Schätzer aus (2.27) die Nichtlinearität $f(x) = \ln|x|$ und erzielt dabei eine sehr gute Sprachsignalqualität. Um von einem möglichen Synergiepotenzial der Nichtlinearitäten $f(x) = \ln|x|$ und $f(x) = |x|^\beta$ zu profitieren, wird hier eine Verkettung der beiden Funktionen vorgeschlagen, die in einer unkonventionellen Nichtlinearität resultiert

$$f(x) = \ln|x|^\beta, \tag{8.2}$$

die soweit dem Autor bekannt bei Herleitung von spektralen Filterfunktionen noch nicht zum Einsatz kam außer im Trivialfall für $\beta = 1$. Man beachte, dass die Nichtlinearität aus (8.2) auch zu $f(x) = \beta \cdot \ln|x|$ umgeformt und als eine einfache Skalierung des natürlichen Logarithmus von $|x|$ betrachtet werden kann. Und obwohl beide Formulierungen dieser Nichtlinearität gleichwertig sind, wird die Fassung (8.2) bevorzugt, da sie intuitiv durch Kombination zweier bekannten Nichtlinearitäten zustande kommt. Denn es scheint nicht ganz klar zu sein, wie eine Skalierung der Nichtlinearität $\ln|x|$ mit einem Vorfaktor β begründet werden soll.

Damit eine solch diffizile Nichtlinearität wie (8.2) zu einer spektralen Filterfunktion führt, welche die Verwendung von speziellen mathematischen Funktionen möglichst vermeidet, wird statt eines MMSE-Schätzers wie in (8.1) ein MAP-basierter Schätzer gesucht, der folgendermaßen definiert werden kann

$$f(\hat{S}(k,\ell)) = \arg\max_z \; p_{Z_\beta(k,\ell)\,|\,Y(k,\ell)}(z\,|\,y), \tag{8.3}$$

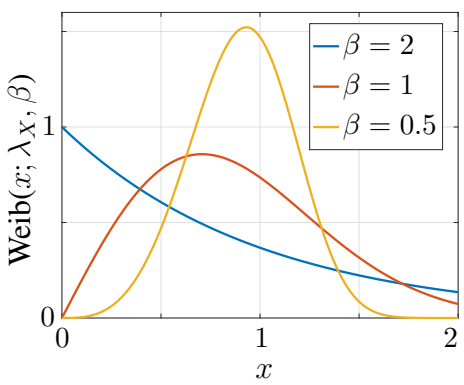
wobei $Z_\beta(k,\ell) = \ln|S(k,\ell)|^\beta$ die logarithmische generalisierte spektrale Amplitude des ungestörten Sprachsignals und z eine Realisierung von $Z_\beta(k,\ell)$ sind. Nachdem die Positionierung der LGSA-Filterfunktion unter den generalisierten Filterfunktionen geklärt und die neue Nichtlinearität eingeführt ist, wird der LGSA-Schätzer in fünf Schritten hergeleitet.

8.1.1. Modellierung der generalisierten spektralen Amplituden

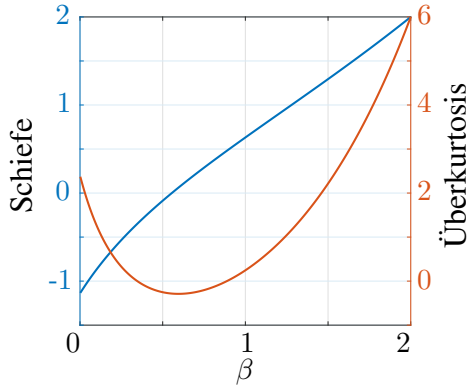
Wie in Abschnitt 2.1 eingeführt wird auch bei den Herleitungen hier angenommen, dass ein ungestörtes Sprachsignal $s(n)$ von einer additiven Störung $d(n)$ im Zeitbereich wie in (2.1) überlagert wird, woraus die Additivität im STFT-Bereich (2.3) resultiert. Die STFT-Koeffizienten $S(k,\ell)$ und $D(k,\ell)$ werden mit (2.4) und (2.5) als unkorrelierte komplexwertige mittelwertfreie normalverteilte Zufallsprozesse modelliert, welche die frequenzabhängigen zeitvarianten Leistungsdichtespektren $\lambda_S(k,\ell)$ und $\lambda_D(k,\ell)$ als Modellparameter haben. In diesem Fall sind die STFT-Koeffizienten des gestörten Sprachsignals $Y(k,\ell)$ auch komplexwertig mittelwertfrei und normalverteilt mit einer Verteilungsdichtefunktion (2.9).

Statistische Modellierung im GSA-Bereich (Schritt 1): Unter diesen Annahmen unterliegen generalisierte spektrale Amplituden $X_\beta(k,\ell)$ aus (3.6) mit $X \in \{Y, S, D\}$ entsprechend einer Zufallsvariablentransformation in den GSA-Bereich den reellwertigen Weibull-Verteilungen

$$p_{X_\beta(k,\ell)}(x) = \mathrm{Weib}(x;\, \lambda_X(k,\ell),\, \beta), \tag{8.4}$$



(a) Weibull-Verteilungen



(b) Schiefe mit Überkurtosis

Abbildung 8.1.: (a) Weibull-Verteilungen für feste Werte des Formparameters $\beta \in \{0.5, 1, 2\}$ und (b) Schiefe und Überkurtosis einer Weibull-Verteilung als Funktionen von $\beta \in [0; 2]$.

welche in [Wei51] eingeführt und ausführlich beschrieben werden. In diesem Kapitel werden die Weibull-Verteilungen etwas anders als in [Wei51] definiert und zwar

$$\text{Weib}(x;\, \lambda_X,\, \beta) \triangleq \frac{2}{\beta \lambda_X} \cdot x^{\frac{2}{\beta}-1} \cdot \exp\left(-\frac{x^{\frac{2}{\beta}}}{\lambda_X}\right) \cdot U(x), \tag{8.5}$$

wobei $U(x)$ die Einheitssprungfunktion (engl. *unit step function*) ist. Besonders erwähnenswert ist die Tatsache, dass die Leistungsdichtespektren $\lambda_X(k, \ell)$ und der Kompressionsfaktor β in die Verteilungsdichtefunktionen der GSA-Koeffizienten jeweils als Skalierungsparameter und Formparameter einfließen. Bei einer Weibull-verteilten Zufallsvariable werden die statistischen Momente der κ-ten Ordnung wie folgt berechnet

$$E[X_\beta^\kappa] = \Gamma\left(\frac{\kappa\beta}{2} + 1\right) \cdot \lambda_X^{\frac{\kappa\beta}{2}}, \tag{8.6}$$

wobei $\Gamma(x)$ eine Gamma-Funktion ist. Da die Weibull-Verteilung (8.5) verschiedene andere Verteilungen mit einschließt, stellt sie sich im Kontext von generalisierten spektralen Filterfunktionen als eine generalisierte Verteilungsdichtefunktion dar. Während eine Exponentialverteilung ein Sonderfall der Weibull-Verteilung für $\beta = 2$ ist, ergibt sich eine Rayleigh-Verteilung für $\beta = 1$, siehe Abb. 8.1 (a). Die Tauglichkeit der Weibull-Verteilung für die statistische Modellierung der spektralen Amplituden wird in [TA10] untersucht.

Zugleich wird die zweiparametrige Weibull-Verteilung selbst von einer generalisierten Gamma-Verteilung aus [EHHJ07] umfasst, die wie folgt definiert werden kann

$$\text{GenGam}(x;\, \lambda, \beta, \eta) \triangleq \frac{2 \cdot U(x)}{\lambda \cdot \beta \cdot \Gamma(\eta)} \cdot x^{\frac{2\eta}{\beta}-1} \cdot \exp\left(-\frac{x^{2/\beta}}{\lambda^{1/\eta}}\right), \tag{8.7}$$

wobei λ, β und η jeweils ein Skalierungsparameter und zwei verschiedene Formparameter sind. Tab. 8.2 verdeutlicht, dass eine Gamma-Verteilung und eine Chi-Verteilung, welche bei den Herleitungen von generalisierten Filterfunktionen auch verwendet werden, zwei weitere Sonderfälle von GGV sind. Während bei diesen Verteilungen der erste Formparameter festgehalten wird, ergibt sich aus einer generalisierten Gamma-Verteilung eine Weibull-Verteilung, wenn man den zweiten Formparameter zu Eins setzt.

Verteilungs-dichtefunktion	Skalierungs-parameter	Form-Parameter 1	Form-parameter 2
Generalisierte Gamma	λ	β	η
Weibull	λ	β	1
Gamma	λ	2	η
Chi	λ	1	η

Tabelle 8.2.: Generalisierte Gamma-Verteilung und ihre zweiparametrige Sonderfälle.

Approximation durch konsistente Normalverteilungen (Schritt 2): Auf dem Weg zu einer recheneffizienten generalisierten spektralen Filterfunktion wird für weitere Berechnungen vorgeschlagen, die Weibull-Verteilungen der GSA-Koeffizienten mit einer reellwertigen Normalverteilung zu approximieren

$$p_{X_\beta(k,\ell)}(x) = \text{Weib}(x;\, \lambda_X,\, \beta) \approx \mathcal{N}\left(x; \mu_X,\, \sigma_X^2\right). \tag{8.8}$$

Dabei sollen die wahre Verteilungsdichtefunktion und ihre Approximation entsprechend der Momentenanpassung (engl. *moment matching approximation*) die gleichen Mittelwerte und Varianzen aufweisen, die unter Verwendung von (8.6) wie folgt berechnet werden

$$\mu_X \triangleq E[X_\beta] = \Gamma\left(\frac{\beta}{2}+1\right)\cdot \lambda_X^{\frac{\beta}{2}} \tag{8.9}$$

$$\sigma_X^2 \triangleq E\left[(X_\beta - \mu_X)^2\right] = c_\beta \cdot \lambda_X^\beta = \frac{c_\beta \cdot \mu_X^2}{\Gamma^2\left(\frac{\beta}{2}+1\right)} \tag{8.10}$$

wobei $c_\beta > 0$ eine positive Konstante ist, die nur vom Kompressionsfaktor β abhängt

$$c_\beta = \Gamma\left(\beta+1\right) - \Gamma^2\left(\frac{\beta}{2}+1\right). \tag{8.11}$$

Man beachte, dass der Vorfaktor $c_\beta/\Gamma^2(\beta/2+1)$ in (8.10) für $\beta > 0$ stets positiv ist und mit dem steigenden Formfaktor β monoton wächst.

Zugegebenermaßen ist der Grund für diese Approximation nicht unmittelbar einleuchtend. Allerdings, wenn man beachtet, dass die Schiefe einer Weibull-Verteilung, die nur vom Formparameter abhängt und in Abb. 8.1 (b) dargestellt ist, für den Wertebereich des Formparameters $\beta \in [0.5;\, 0.6]$ etwa Null ist, scheint die vorgeschlagene Approximation zumindest hier durchaus sinnvoll zu sein. So sieht die Weibull-Verteilung für $\beta = 0.5$ einer Normalverteilung sehr ähnlich aus, siehe Abb. 8.1 (a). Eine weitere Rechtfertigung der Approximation (8.8) wird in Unterabschnitt 8.1.2 gegeben. Sonst soll erwähnt werden, dass eine Normalverteilung mit den Parametern wie in (8.9) und (8.10) besondere Eigenschaften hat, denn ihr Mittelwert (8.9) ist stets positiv und ihre Varianz (8.10) ist kein von μ_X unabhängiger Parameter, sondern weist für $\beta > 0$ einen monoton steigenden Zusammenhang als Funktion von μ_X auf. Normalverteilungen mit solchen Eigenschaften werden beispielsweise in [RSU00] als konsistente Normalverteilungen bezeichnet.

8.1.2. Alternative Herleitung des MMSE-Schätzers

Der Vorteil der Approximation durch Normalverteilungen besteht darin, dass man im dritten Herleitungsschritt einen MMSE-Schätzer im GSA-Bereich finden kann, die sich als ein bereits bekannter generalisierter Schätzer aus [STCT98] darstellt.

Berechnung einer bedingten Verteilungsdichtefunktion im GSA-Bereich (Schritt 3): Um die bedingte Verteilungsdichtefunktion $p_{S_\beta|Y_\beta}(s|y)$ zu berechnen, werden zum Einen die Annahme der Additivität der generalisierten spektralen Amplituden (3.1) und zum Anderen die konsistenten Normalverteilungen (8.8) benötigt. Mit dem Satz von Bayes für bedingte Wahrscheinlichkeitsdichtefunktionen ergibt sich dann für $p_{S_\beta|Y_\beta}(s|y)$ eine Normalverteilung

$$p_{S_\beta|Y_\beta}(s|y) = \frac{p_{D_\beta}(y-s) \cdot p_{S_\beta}(s)}{p_{Y_\beta}(y)} = \mathcal{N}(s; \mu_{S|Y},\, \sigma^2_{S|Y}), \tag{8.12}$$

deren Mittelwert $\mu_{S|Y}$ und Varianz $\sigma^2_{S|Y}$ aus den Parametern der Normalverteilungen (8.8) mit (8.9) und (8.10) berechnet werden können. Dabei stellt sich heraus, dass der Mittelwert $\mu_{S|Y} = E\,[\,S_\beta \,|\, Y_\beta\,]$ dem MMSE-Schätzer der parametrischen generalisierten spektralen Subtraktion $\hat{S}_\beta$ aus (3.8) entspricht, der in [STCT98] mit einer anderen Herangehensweise bereits hergeleitet wurde. Wie in Abschnitt 3.2 beschrieben, wird der MMSE-PGSS-Schätzer zunächst so definiert, dass er eine spektrale Subtraktion im GSA-Bereich unter Verwendung von zwei zu optimierenden Parametern realisiert. Anschließend werden diese Parameter so gewählt, dass sie den MSE-Fehler $E[(S_\beta - \hat{S}_\beta)^2]$ minimieren, wobei die spektralen Amplituden der beteiligten Zufallsprozesse mit Hilfe einer Rayleigh-Verteilung modelliert werden, siehe Tab. 8.1. Die spektrale Filterfunktion des MMSE-PGSS-Schätzers G_β^{PGSS} aus (3.9) hängt dabei unter anderem vom Kompressionsfaktor β ab, der in [STCT98] für experimentelle Untersuchungen zu den Werten $\beta = 1$ und $\beta = 2$ gesetzt wird. Die Übereinkunft von $\mu_{S|Y}$ mit dem MMSE-PGSS-Schätzer kann als eine willkommene Rechtfertigung der Approximationen (3.1) und (8.8) betrachtet werden.

Ein großer Vorteil der vorgeschlagenen statistischen Modellierung gegenüber der Herangehensweise in [STCT98] ist allerdings die Möglichkeit, neben dem MMSE-Schätzer $\mu_{S|Y}$ auch die Varianz $\sigma^2_{S|Y}$ der bedingten Verteilung (8.12) analytisch zu berechnen. Für die weiteren Berechnungen werden die beiden Statistiken gemeinsam angegeben:

$$\mu_{S|Y} = G_\beta^{\text{PGSS}} \cdot Y_\beta, \tag{8.13}$$

$$\sigma^2_{S|Y} = \frac{c_\beta}{\gamma^\beta} \cdot \frac{\xi^\beta}{\xi^\beta + 1} \cdot Y_\beta^2, \tag{8.14}$$

wobei γ und ξ jeweils das *a posteriori* SNR aus (2.16) und das *a priori* SNR aus (2.22) sind. Man beachte, dass in (8.13) und (8.14) auf die Indizierung (k, ℓ) bei den jeweiligen Größen der Übersichtlichkeit halber verzichtet wird. Die berechnete bedingte Verteilungsdichtefunktion (8.12) mit bekannten Parametern (8.13) und (8.14) kann für eine Zufallsvariablentransformation verwendet werden, die einen Übergang vom GSA-Bereich in den Bereich der logarithmischen GSA-Koeffizienten realisiert.

8.1.3. MAP-basierter Schätzer der logarithmischen GSA

Um die gesuchte Verteilungsdichtefunktion im Bereich der logarithmischen GSA zu berechnen, fehlen jetzt nur noch zwei Herleitungsschritte, die einen Übergang in den Bereich der generalisierten log-spektralen Amplituden realisieren.

Trunkierung der bedingten VDF (Schritt 4): Als Vorbereitung auf die bevorstehende Zufallsvariablentransformation muss die Verteilungsdichtefunktion $p_{S_\beta|Y_\beta}(s|y)$ aus (8.12) zunächst auf den Bereich der positiven Argumente $s > 0$ begrenzt werden, da alle Realisierungen von S_β per definitionem (3.6) positiv sind. Somit kann S_β als eine gestutzte (engl. *truncated*) Zufallsvariable betrachtet werden. Basierend auf diesen Überlegungen wird vorgeschlagen, die VDF $p_{S_\beta|Y_\beta}(s|y)$ durch eine bei $s = 0$ abgeschnittene rechtsseitige Verteilung mit demselben Modalwert zu ersetzen

$$\tilde{p}_{S_\beta|Y_\beta}(s|y) = \frac{U(s)}{\Phi\left(\frac{\mu_{S|Y}}{\sigma_{S|Y}}\right)} \cdot \mathcal{N}(s; \mu_{S|Y},\, \sigma^2_{S|Y}), \tag{8.15}$$

wobei $\Phi(x)$ die Verteilungsfunktion einer Standardnormalverteilung $\mathcal{N}(x;\, 0,\, 1)$ ist [PP02].

Berechnung der VDF im LGSA-Bereich (Schritt 5): Ausgehend von $\tilde{p}_{S_\beta|Y_\beta}(s|y)$ kann jetzt mit der Zufallsvariablentransformation die bedingte VDF $p_{Z_\beta|Y_\beta}(z|y)$ der generalisierten GSA $Z_\beta = \ln S_\beta$ berechnet werden:

$$p_{Z_\beta|Y_\beta}(z|y) = \frac{e^z \cdot \mathcal{N}(e^z; \mu_{S|Y},\, \sigma^2_{S|Y})}{\Phi\left(\frac{\mu_{S|Y}}{\sigma_{S|Y}}\right)} \;\propto\; e^{\frac{1}{2}\cdot f(z)} \tag{8.16}$$

wobei $f(z) = 2\cdot z - \left(e^z - \mu_{S|Y}\right)^2 / \sigma^2_{S|Y}$ eine Hilfsfunktion ist, die in weiteren Berechnungen verwendet wird. Da die Herleitung eines MMSE-Schätzers im LGSA-Bereich zu speziellen mathematischen Funktionen führt, wird hier vorgeschlagen, einen einfachen MAP-Schätzer der LGSA-Koeffizienten aufzustellen, der mit (8.3) und (8.16) über die Hilfsfunktion gefunden werden kann

$$\hat{Z}_\beta = \arg\max_z \; f(z) = z_{\max}. \tag{8.17}$$

Die Maximumstelle $z_{\max}$ von $f(z)$ entspricht der Nullstelle der ersten Ableitung von $f(z)$ und kann über die Gleichung $e^{2z} - \mu_{S|Y} \cdot e^z - \sigma^2_{S|Y}\Big|_{z=z_{\max}} \overset{!}{=} 0$ bestimmt werden, welche sich im GSA-Bereich trivial lösen lässt. Die einzige positive Nullstelle dieser Gleichung im GSA-Bereich ist der gesuchte MAP-LGSA-Schätzer:

$$\hat{S}^{\text{LGSA}}_\beta = \frac{\mu_{S|Y}}{2} + \sqrt{\left(\frac{\mu_{S|Y}}{2}\right)^2 + \sigma^2_{S|Y}}. \tag{8.18}$$

Verwendung von (8.13) und (8.14) in (8.18) führt zur spektralen MAP-LGSA-Filterfunktion

$$G_{\text{LGSA}}(k,\ell) = [G^{\text{LGSA}}_\beta(k,\ell)]^{\frac{1}{\beta}}, \tag{8.19}$$

wobei die Filterfunktion im GSA-Bereich mit (3.9) wie folgt ausgeschrieben werden kann

$$G^{\text{LGSA}}_\beta(k,\ell) = \frac{G^{\text{PGSS}}_\beta}{2} + \sqrt{\left(\frac{G^{\text{PGSS}}_\beta}{2}\right)^2 + \frac{c_\beta}{\gamma^\beta} \cdot \frac{\xi^\beta}{\xi^\beta + 1}}. \tag{8.20}$$

Der MAP-LGSA-Schätzer $\hat{S}^{\text{LGSA}}$ steht aufgrund der Ungleichung $G^{\text{LGSA}}_\beta > G^{\text{PGSS}}_\beta$, die für einen bestimmten Wert von β immer gilt, in der Relation $\hat{S}^{\text{LGSA}} > \hat{S}^{\text{PGSS}}$ zum MMSE-PGSS-Schätzer. Soweit dem Autor bekannt, ist die spektrale Filterfunktion (8.20) die erste MAP-basierte Filterfunktion im log-spektralen Bereich, siehe Tab. 3.2 oder vergleiche mit [LL11].

Man beachte, dass die Gamma-Funktion, welche bei der Berechnung von G_β^{PGSS} in (3.9) und von c_β in (8.11) vorkommt, für einen bestimmten Wert von β zu einer Konstanten wird, sodass die hergeleitete Filterfunktion in einer Realisierung in der Tat von keinen speziellen mathematischen Funktionen Gebrauch macht und somit wie gewünscht recheneffizient implementiert werden kann.

8.2. Erhöhung der Modellflexibilität und Parameteroptimierung

Inspiriert durch die Aussage von George E. P. Box '*Classical methods of estimation should be retained using models which more appropriately represent reality*' aus [Box79] wird im Folgenden eine Strategie vorgeschlagen, wie man die Leistungsfähigkeit des verwendeten Schätzers hinsichtlich der Sprachsignalqualität steigern kann. Dabei wird ähnlich wie in [CHHU16] die Flexibilität der verwendeten statistischen Modelle erhöht. Und da der MMSE-PGSS-Schätzer bei Herleitung des MAP-LGSA-Schätzers als Nebenprodukt entsteht, profitiert er auch von der erhöhten Modellflexibilität, woraus ein modifizierter MMSE-PGSS-Schätzer resultiert.

Erhöhung der Modellflexibilität: Die Leistungsfähigkeit der LGSA-Filterfunktion soll dadurch verbessert werden, dass die Verteilungsdichtefunktionen (9.2) der beteiligten Zufallsprozesse einen unabhängigen Formparameter $\vartheta \in \mathbb{R}_{>0}$ bekommen, welcher nicht zwangsläufig gleich dem Kompressionsfaktor β ist, der bei der Berechnung von generalisierten spektralen Amplituden in (3.6) verwendet wird. Also wird die statistische Modellierung (9.2) folgendermaßen modifiziert

$$p_{X_\beta(k,\ell)}(x) = \text{Weib}(x;\, \lambda_X(k,\ell),\, \vartheta). \tag{8.21}$$

Diese Modifikation gibt den statistischen Modellen eine Möglichkeit, sich an die Eigenschaften der Sprachdaten anzupassen, die in der Realität den getroffenen Annahmen nicht gerecht werden, wie z. B. Missachtung der Dünnbesetztheit spektraler Darstellung von Sprachsignalen. Die Fünf-Schritte-Herleitung der LGSA-Filterfunktion aus Abschnitt 8.1 unter Verwendung von (8.21) statt (9.2) führt zu den Filterfunktionen im GSA-Bereich aus (3.9) und (8.20), die nicht mehr von β sondern von ϑ abhängen. Die auf diese Weise modifizierte PGSS-Filterfunktion aus (3.9) ergibt sich dann zu

$$G_{\text{PGSS}} = \left[G_\vartheta^{\text{PGSS}}\right]^{\frac{1}{\beta}} = \left[\frac{\xi^\vartheta}{\xi^\vartheta + 1} \cdot \left(1 - \frac{\Gamma\left(\frac{\vartheta}{2}+1\right)}{\gamma^{\frac{\vartheta}{2}}}\left(1 - \xi^{-\frac{\vartheta}{2}}\right)\right)\right]^{\frac{1}{\beta}}. \tag{8.22}$$

Diese kann für eine Entstörung spektraler Amplituden entsprechend (2.17) verwendet werden. Außerdem resultiert für die LGSA-Filterfunktion folgende Berechnungsvorschrift:

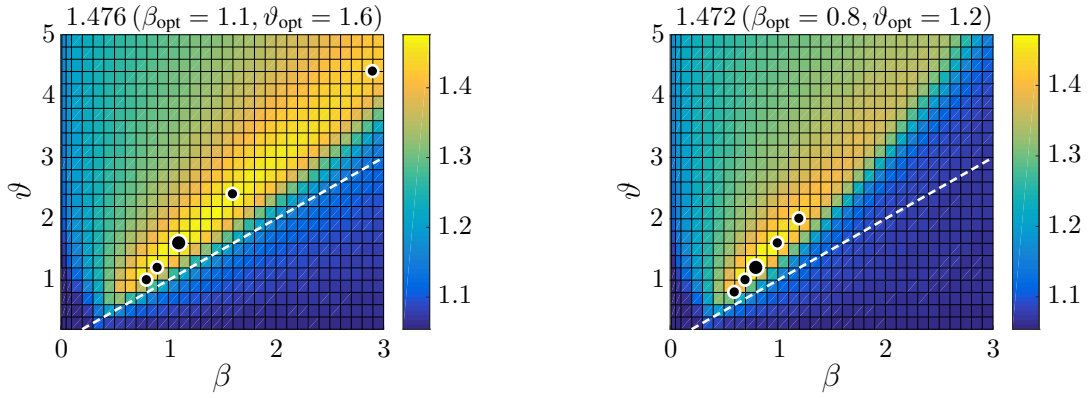
$$G_{\text{LGSA}} = \left[\frac{G_\vartheta^{\text{PGSS}}}{2} + \sqrt{\left(\frac{G_\vartheta^{\text{PGSS}}}{2}\right)^2 + \frac{c_\vartheta}{\gamma^\vartheta} \cdot \frac{\xi^\vartheta}{\xi^\vartheta + 1}}\right]^{\frac{1}{\beta}}. \tag{8.23}$$

Nun sollen die Parameter ϑ und β in den Experimenten mit den Sprachsignalen der TIMIT Datenbank, die additiv mit dem weißen Rauschen gestört sind, so ermittelt werden, dass

sie eingesetzt in einem System zur spektralen Sprachsignalentstörung zu den prozessierten Signalen mit bester Sprachsignalqualität gemessen in MOS-LQO$_{\text{WB}}$ führen.

Datengetriebene Optimierung der Parameter β und ϑ: Für die bevorstehende Parameteroptimierung werden die ungestörten Sprachsignalaufnahmen der TIMIT-Datenbank weiblicher und männlicher Sprecher jeweils zu den einminütigen Signalen zusammengesetzt. Diese werden anschließend mit einem weißen Störsignal der NOISEX-92 Datenbank so additiv überlagert, dass die gestörten Sprachsignale ein eingangsseitiges globales SNR$_{\text{IN}}$ von $\{-5, 0, 5, 10, 15\}$ dB aufweisen. Dabei sind alle Signale mit einer Abtastrate von 16 kHz abgetastet. Man beachte, dass die beiden verwendeten Datenbanken in Abschnitt 2.4 bereits beschrieben wurden. Für die STFT aus Abschnitt 2.1 werden ein Hamming-Analysefenster der FFT-Länge von $K = 512$ Abtastwerten mit einem Rahmenvorschub von $R = 128$ Abtastwerten verwendet. Während das konventionelle *Minimum Statistics* Verfahren mit denselben Parametern wie in [Mar01] für die Schätzung der Rauschleistungsdichte $\lambda_D(k, \ell)$ eingesetzt wird, wird das *a priori* SNR $\xi(k, \ell)$ mit dem *Decision-Directed* Verfahren aus [EM84] geschätzt, dessen Gewichtsparameter zu $\alpha_{\text{DD}} = 0.975$ mit einem minimalen Wert des *a priori* SNR von $\xi_{\text{min}} = -25\,\text{dB}$ wie in [Coh04] gesetzt wird. Sonst wird $G_{\text{min}} = -25\,\text{dB}$ als Untergrenze der Filterfunktionen verwendet wie in [Coh01]. Man beachte, dass im verwendeten System zur spektralen Sprachsignalentstörung auf den Baustein 4 aus Abb. 2.2 verzichtet wird. Die Qualität der prozessierten Sprachsignale wird mit breitbandigen MOS-LQO-Werten gemessen, die in Abschnitt 2.3 bereits beschrieben wurden. Die zu optimierenden Parameter β und ϑ werden in dieser Untersuchung jeweils in den Wertebereichen $[0.01;\ 3]$ und $[0.2;\ 5]$ variiert.

In Abb. 8.2 sind resultierende MOS-LQO-Werte gemittelt über die Signalen mit weiblichen und männlichen Sprechern für SNR$_{\text{IN}} = 5\,\text{dB}$ dargestellt und zwar in Abb. 8.2 (a) für die modifizierte PGSS-Filterfunktion aus (8.22) und in Abb. 8.2 (b) für die vorgeschlagene LGSA-Filterfunktion aus (8.23). In den Überschriften der Bilder sind neben den maximal erreichten Werten MOS-LQO$_{\text{max}}$ auch die dazugehörigen optimalen Parameter $(\beta_{\text{opt}},\ \vartheta_{\text{opt}})$



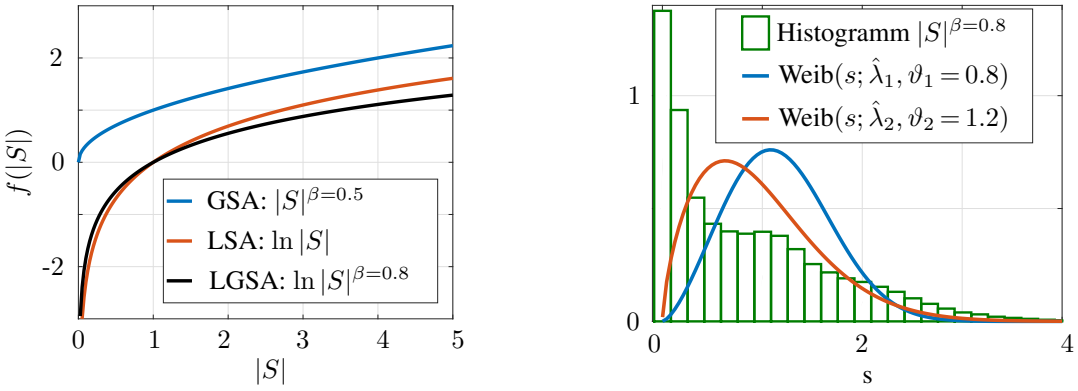
(a) PGSS-Filterfunktion (b) LGSA-Filterfunktion

Abbildung 8.2.: Durchschnittliche MOS-LQO-Werte in den Experimenten mit weißem Rauschen beim SNR$_{\text{IN}} = 5\,\text{dB}$ mit optimalen Parametern (großer schwarzer Punkt): (a) PGSS-Filterfunktion aus (8.22), (b) LGSA-Filterfunktion aus (8.23). Kleinere schwarze Punkte entsprechen den optimalen Parametern für $\{-5, 0, 10, 15\}$ dB, wobei kleinere Werte der Parameter mit höheren SNR$_{\text{IN}}$-Werten einhergehen.

angegeben, die in den Bildern mit den großen schwarzen Punkten gekennzeichnet sind. Die experimentelle Untersuchung zeigt, dass die beiden Filterfunktionen bei optimaler Parametrisierung in Gegenwart eines weißen Störsignals eine identische Qualität der prozessierten Sprachsignale erreichen können. Die optimalen Werte $(\beta_{\text{opt}}, \vartheta_{\text{opt}})$ unterscheiden sich allerdings, sodass die PGSS-Filterfunktion mit $\beta_{\text{opt}} = 1.1$ und $\vartheta_{\text{opt}} = 1.6$ etwas höhere Parameterwerte erfordert als die LGSA-Filterfunktion mit $\beta_{\text{opt}} = 0.8$ und $\vartheta_{\text{opt}} = 1.2$. Außerdem fällt auf, dass der MMSE-basierter PGSS-Schätzer gute MOS-LQO-Werte in einem etwas größeren Wertebereich der zu optimierenden Parameter liefert als der MAP-basierter LGSA-Schätzer. Allerdings, wenn man die SNR_{IN}-Werte der Eingangssignale variiert, streuen die optimalen Punkte, die in Abb. 8.2 als kleine schwarze Punkte zu sehen sind, beim modifizierten PGSS-Schätzer deutlich stärker als beim LGSA-Schätzer, was bei einer nichtadaptiven Parametrisierung der Filterfunktionen mit festen Parameterwerten nachteilig ist. Dabei gehen die kleineren Parameterwerte der optimalen Punkte mit höheren SNR_{IN}-Werten einher und umgekehrt.

Interessanterweise liegen alle Optima der jeweiligen Filterfunktion auf einer Geraden, die in beiden Fällen eine Neigung größer Eins aufweist, sodass keiner der optimalen Punkte auf der Geraden $\vartheta = \beta$ liegt, die in Abb. 8.2 als eine gestrichelte weiße Gerade dargestellt wird. Somit wird Nützlichkeit der vorgeschlagenen Erhöhung der Flexibilität der statistischen Modelle durch Entkopplung des Formparameters einer Weibull-Verteilung ϑ vom Kompressionsfaktors β nachgewiesen. Außerdem wird die Stellungnahme aus [ITS$^+$10] bestätigt, dass der Bereich der spektralen Amplituden $(\beta = 1)$ oder der spektralen Leistungen $(\beta = 2)$ nicht immer optimal hinsichtlich der besten Qualität von entstörten Sprachsignalen ist. Obwohl die vorliegende Auswertung nahelegt, dass man die optimalen Parameter der beiden Filterfunktionen adaptiv in Abhängigkeit vom SNR_{IN}-wählen könnte, wird in weiteren Experimenten darauf verzichtet, denn sonst müsste man den SNR_{IN}-Wert zusätzlich schätzen. Also werden für die weiteren Analysen und experimentelle Untersuchungen die nichtadaptiven optimalen Parameter verwendet, die sich für $\text{SNR}_{\text{IN}} = 5\,\text{dB}$ ergaben.

Vergleich der Nichtlinearitäten und Einfluss der Parameterentkopplung: Die optimalen Parameterwerte β_{opt} und ϑ_{opt} der LGSA-Filterfunktion lassen Rückschlüsse auf resultierende Nichtlinearität und auf bevorzugte statistische Modellierung von Daten ziehen, die laut Experimenten zur besten Qualität der prozessierten Sprachsignale führen. Da die vorgeschlagene Nichtlinearität (8.2) sich als eine Kombination von solchen etablierten Nichtlinearitäten wie die Kompressionsfunktion und der natürliche Logarithmus darstellt, werden diese drei Nichtlinearitäten in Abb. 8.3 (a) dargestellt. Dabei wird der Kompressionsfaktor der Kompressionsfunktion auf den Wert $\beta = 0.5$ gesetzt, der laut den Untersuchungen in [BKM08] in einer guten Störsignaldämpfung ohne erkennbaren Verzerrungen des Sprachsignals resultiert und in einer Wurzel-Kompression der spektralen Amplituden mündet, die im GSA-Bereich angesiedelt ist. Man beachte, dass die Wurzel-Kompression auch in den Untersuchungen in [ITS$^+$10] verwendet wird. Außerdem wird in Abb. 8.3 (a) der natürliche Logarithmus der spektralen Amplituden dargestellt, welcher bei der Entwicklung der MMSE-LSA-Filterfunktion aus [EM85] verwendet wird, die für eine gute Qualität der prozessierten Sprachsignale bekannt ist. Während die spektralen Amplituden durch die Wurzel-Kompression betragsmäßig auf den Bereich der positiven GSA-Koeffizienten komprimiert werden, werden sie durch die Logarithmus-Funktion für Werte größer Eins zwar auch komprimiert jedoch für die kleineren Werte auf den ganzen Wertebereich der negativen LSA-Koeffizienten auseinander gezogen. Unterschiedliche Wirkungsweisen dieser beiden Nicht-



(a) Vergleich der Nichtlinearitäten (b) Einfluss der Parameterentkopplung

Abbildung 8.3.: (a) Vergleich der Nichtlinearitäten und (b) Einfluss der Entkoppelung des Formfaktors ϑ einer Weibull-Verteilung vom Kompressionsfaktor β auf die statistische Modellierung von Sprachdaten.

linearitäten auf die spektralen Amplituden könnten verschiedene Eigenschaften der dazugehörigen spektralen Filterfunktion hinsichtlich der Sprachsignalentstörung erklären. Die vorgeschlagene Nichtlinearität (8.2) wird in Abb. 8.3 (a) für den optimalen Wert der LGSA-Filterfunktion von $\beta_{\text{opt}} = 0.8$ aus Abb. 8.2 (b) dargestellt. Wie man sieht, weicht die neue Nichtlinearität, die auch in der Form $\beta \cdot \ln|S|$ geschrieben werden kann, nicht all zu sehr von dem natürlichen Logarithmus ab. Nennenswerte Abweichungen sind allerdings bei größeren Werten der spektralen Amplituden zu beobachten, wo die vorgeschlagene Nichtlinearität interessanterweise in etwa die Steigung der Kompressionsfunktion aus [BKM08] aufweist. Welche Auswirkung die neue Nichtlinearität auf die Eigenschaften der entstörten Sprachsignale hat, werden weitere Analysen und experimentelle Untersuchungen in Abschnitt 8.3 zeigen.

Davor soll allerdings der Einfluss der durchgeführten Optimierung auf die statistische Modellierung der GSA-Koeffizienten beim LGSA-Schätzer diskutiert werden. Wie bereits erwähnt, erzielt die LGSA-Filterfunktion die beste Sprachsignalqualität, wenn der Formparameter der Weibull-Verteilung ϑ größer als der dazugehörige Kompressionsfaktor β gewählt wird. Dieser Sachverhalt ist in Abb. 8.3 (b) verdeutlicht, in der ein normiertes Histogramm der GSA-Koeffizienten eines ungestörten Sprachsignals $|S|^{\beta}$ für $\beta = 0.8$ mit zwei Weibull-Verteilungen zu sehen ist, deren Skalierungsparameter λ mit der Maximum-Likelihood Methode geschätzt werden. Viele energieschwache und wenige energiereiche Beobachtungen des normierten Histogramms weisen darauf hin, dass die GSA-Koeffizienten eines ungestörten Sprachsignals dünnbesetzt sind. Beim Betrachten der beiden Weibull-Verteilungen mit verschiedenen Formparametern $\vartheta_1 = 0.8$ und $\vartheta_2 = 1.2$ ist zu beachten, dass die Kurtosis einer Weibull-Verteilung laut [Rou73] ähnlich wie seine Schiefe unabhängig vom Skalierungsparameter ist und für die Werte der Formparameters größer als 0.6 mit steigenden Werten von ϑ wächst, siehe Abb. 8.1 (b). Man beachte, dass in Unterabschnitt 8.1.1 der Formparameter einer Weibull-Verteilung gleich dem Kompressionsfaktor β gesetzt wurde, sodass in Abb. 8.1 (b) sowohl die Schiefe als auch die Überkurtosis als Funktionen von β dargestellt sind. Somit werden beim LGSA-Schätzer die generalisierten spektralen Amplituden mit den Verteilungen modelliert, welche eine höhere Kurtosis aufweisen, als dies für

$\vartheta = \beta$ der Fall ist. Berücksichtigt man, dass die Überkurtosis (engl. *excess kurtosis*) einer Weibull-Verteilung laut Abb. 8.1 (b) für den Formparameter $\vartheta > 0.89$ positiv ist, resultiert daraus, das die durchgeführte Optimierung mit $\vartheta_{\text{opt}} = 1.2$ zur statistischen Modellierung der beteiligten GSA-Koeffizienten mit steilgipfligen Verteilungen führt. Wie in Abschnitt 3.2 erwähnt, werden bei Herleitung generalisierter spektraler Filterfunktionen steilgipflige Verteilungen durchaus verwendet, da durch diese Dünnbesetztheit von spektralen Koeffizienten eines Sprachsignals statistisch besser modelliert werden kann.

Resultierende Eigenschaften der optimierten Filterfunktion: Außerdem lohnt es sich, Verläufe der vorgeschlagenen LGSA-Filterfunktion mit denen der PGSS- und LSA-Filterfunktionen zu vergleichen. Als Funktionen von zwei Parametern, des *a posteriori* SNR γ und des *a priori* SNR ξ, können spektrale Filterfunktionen ähnlich wie in [EM84, EM85] für verschiedene feste Werte des *a priori* SNR ξ über ein sogenanntes momentanes SNR aufgetragen werden, das als $\gamma - 1$ für $\gamma > 1$ definiert wird, siehe Abb. 8.4. Dabei werden PGSS- und LGSA-Filterfunktionen mit den optimalen Parametern $(\beta_{\text{opt}}; \vartheta_{\text{opt}})$ aus Abb. 8.2 parametrisiert, die für das globale eingangsseitige $\text{SNR}_{\text{IN}} = 5\,\text{dB}$ die beste Sprachsignalqualität liefern. Betrachtet man zunächst die Verläufe der LSA-Filterfunktion, fällt auf, dass sie für alle betrachteten Werte von ξ mit steigendem *a posteriori* SNR sinken. Laut [Cap94, STCT98] können Filterfunktionen mit solchen Eigenschaften das musikalische Rauschen erfolgreich bekämpfen und somit für prozessierte Sprachsignale mit guter Sprachsignalqualität sorgen. Im Unterschied zum LSA-Schätzer weist die PGSS-Filterfunktion eine solche Eigenschaft nur für kleine Werte des *a priori* SNR auf. Für $\xi = 5\,\text{dB}$ ist die PGSS-Filterfunktion monoton steigend und führt somit zu einer stärkeren Störsignaldämpfung für kleinere Werte des *a posteriori* SNR, die auf Kosten der Sprachsignalqualität erreicht wird. Begutachtet man die dargestellten Verläufe der LGSA-Filterfunktion, fällt auf, dass die vorgeschlagene neue Nichtlinearität (8.1) und die durchgeführte Parameteroptimierung zu den Funktionsverläufen führen, die sich in etwa zwischen den jeweiligen Kurven der LSA- und PGSS-Filterfunktionen platzieren und somit versuchen, die Vorteile der beiden Schätzer auszunutzen. Die LGSA-Filterfunktion versucht somit, die gewünschte Eingenschaft des LSA-Schätzers hinsichtlich einer guten Qualität der prozessierten Signale zu behalten und dabei gleichzeitig eine bes-

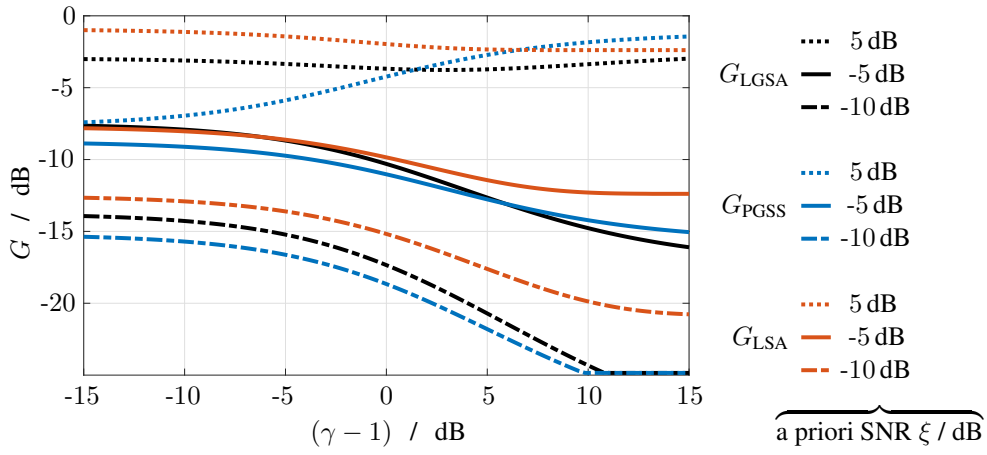


Abbildung 8.4.: Filterfunktionen der LGSA-, PGSS- und LSA-Schätzer im Vergleich untereinander für feste Werte des *a priori* SNR $\xi \in \{\text{-}10, \text{-}5, 5\}$dB als Funktionen des momentanen *a posteriori* SNR $(\gamma - 1)$ für das *a posteriori* SNR $\gamma > 1$.

sere Störsignaldämpfung zu erreichen. Nun sollen die angestellten Mutmaßungen hinsichtlich der Eigenschaften der betrachteten Filterfunktionen in experimentellen Untersuchungen überprüft werden.

8.3. Leistungsfähigkeit der neuen Filterfunktion auf CHiME-3-Daten

Um Leistungsfähigkeit der betrachteten Filterfunktionen auf den Daten zu untersuchen, die während der Optimierung der PGSS- und LGSA-Filterfunktionen in Abschnitt 8.2 nicht verwendet wurden, wird eine spektrale Sprachsignalentstörung auf den CHiME-3-Daten durchgeführt, die in Abschnitt 2.4 bereits beschrieben wurden. Dabei wird dasselbe System zur spektralen Entstörung verwendet, das in den Experimenten in Abschnitt 8.2 beschrieben wurde. Die PGSS- und LGSA-Filterfunktionen werden dabei wieder mit den nichtadaptiven optimalen Parametern $(\beta_{\text{opt}}; \vartheta_{\text{opt}})$ parametrisiert, die für das globale eingangsseitige $\text{SNR}_{\text{IN}} = 5\,\text{dB}$ die beste Sprachsignalqualität liefern und in den Filterfunktionen resultieren, die in Abb. 8.4 präsentiert wurden. Die Leistungsfähigkeit der Filterfunktionen hinsichtlich der Sprachsignalqualität wird mit Hilfe der Erhöhung des breitbandigen $\Delta\text{MOS-LQO}_{\text{WB}}$-Maßes im Bezug auf das Eingangssignal bewertet

$$\Delta\text{MOS-LQO}_{\text{WB}} = \text{MOS-LQO}_{\text{WB,OUT}} - \text{MOS-LQO}_{\text{WB,IN}}, \tag{8.24}$$

wobei $\text{MOS-LQO}_{\text{WB,OUT}}$ und $\text{MOS-LQO}_{\text{WB,IN}}$ jeweils die breitbandigen MOS-LQO-Werte gemessen am Ausgang und am Eingang des Systems sind. Hinsichtlich Bewertung der Störsignaldämpfung wird die Verbesserung des globalen SNR-Wertes ΔSNR aus (2.41) verwendet.

In Abb. 8.5 sind die mittleren $\Delta\text{MOS-LQO}_{\text{WB}}$-Werte über die durchschnittlichen ΔSNR Werte für jede der vier Störumgebungen der CHiME-3-Daten aufgetragen. Außerdem werden für jede Filterfunktion die Bewertungsmaße gemittelt über alle Störungsarten (avg) mit Pentagrammen separat dargestellt. Wie die Ergebnisse zeigen, erreichen verschiedene Filterfunktionen durchaus unterschiedliche Leistungsfähigkeit in der Sprachsignalentstörung. Dabei gelingt es ihnen, sowohl die Qualität der Eingangssignale zu verbessern als auch ihre Entstörung zu bewirken. Wie erwartet, erreicht die LSA-Filterfunktion im Vergleich zu beiden generalisierten Schätzern die größte Verbesserung in der Sprachsignalqualität. Allerdings geschieht dies auf Kosten einer reduzierten Störsignaldämpfung, wie die Verläufe der Filterfunktion in Abb. 8.4 vermuten ließen. Im Gegensatz dazu erzielt die PGSS-Filterfunktion aus (8.22) eine viel bessere Störsignaldämfung als der LSA-Schätzer, kann jedoch hinsichtlich der Sprachsignalqualität mit dem letzten nicht mithalten. Ein Vergleich der Ergebnisse für die LSA- und PGSS-Filterfunktionen könnte auf den bekannten Kompromiss zwischen der Sprachsignalqualität und der Störsignaldämpfung hindeuten, der die Leistungsfähigkeit der konventionellen modellbasierten Ansätze zur spektralen Sprachsignalentstörung begrenzt. Allerdings gelingt es dem vorgeschlagenen LGSA-Schätzer laut den Experimenten auf CHiME-3-Daten, diesen Kompromiss etwas besser zu lösen und zwar so, dass die Störsignaldämpfung des PGSS-Schätzers doch noch weiter ausgebaut werden kann und dabei keine nennenswerten Verluste der Sprachsignalqualität im Vergleich zum LSA-Schätzer entstehen. Verglichen mit dem LSA-Schätzer verbessert die vorgeschlagene

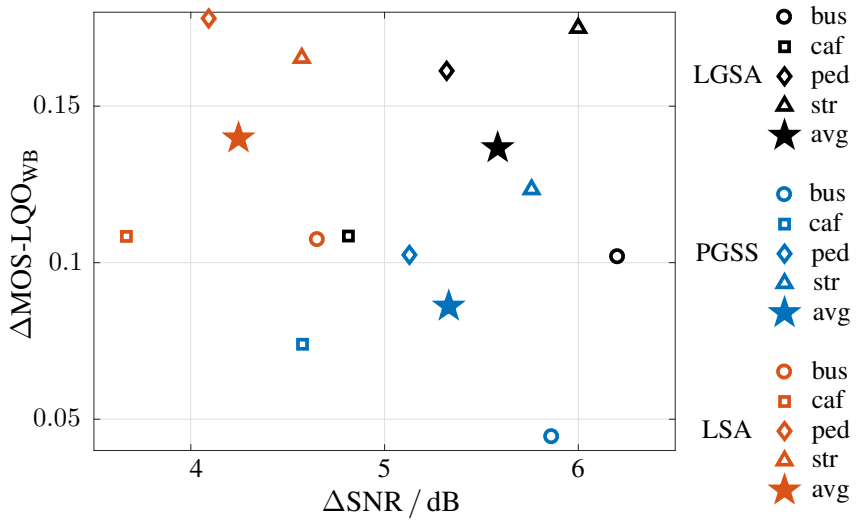


Abbildung 8.5.: Mittlere Verbesserung der Sprachsignalqualität und der Störsignaldämpfung gemessen jeweils in ΔMOS-LQO$_{\text{WB}}$- und ΔSNR-Werten auf den CHiME-3-Daten beim Einsatz der LGSA-, PGSS- und LSA-Filterfunktionen in vier verschiedenen Störumgebungen. Eine Mittelung über alle Störumgebungen (avg) sind mit Pentagrammen angegeben.

LGSA-Filterfunktion die Störsignaldämpfung von 4.2 dB auf 5.6 dB, also im Mittel um etwa 1.4 dB. Die höhere Leistungsfähigkeit ist dabei zum Einen auf die Verwendung der neuartigen Nichtlinearität und zum Anderen auf eine datengetriebene Optimierung zurückzuführen.

Betrachtet man jede der vier Störumgebungen für sich, stellt man fest, dass die Entstörung von Signalen aufgenommen in einer Cafeteria (caf), eine besondere Herausforderung für alle drei Schätzer darstellten. Im Gegensatz dazu ließen sich die Aufnahmen auf einer Straße (str) von allen Filterfunktionen etwas leichter entstören. Außerdem fällt auf, dass der vorgeschlagene LGSA-Schätzer in dieser Störumgebung die LSA-Filterfunktion sowohl hinsichtlich der Sprachsignalqualität als auch der Störsignaldämpfung übertrifft.

8.4. Zusammenfassung

In diesem Kapitel wurde eine spektrale Filterfunktion mit Hilfe einer neuartigen Nichtlinearität hergeleitet, die im Bereich der logarithmischen generalisierten spektralen Amplituden angesiedelt ist. Dabei stellte sich die LGSA-Nichtlinearität als eine Kombination zweier weit verbreiteter Nichtlinearitäten dar, eines natürlichen Logarithmus und einer Kompressionsfunktion, welche in der generalisierten modellbasierten Sprachsignalentstörung häufig zum Einsatz kommt. Um eine einfache recheneffiziente Berechnungsvorschrift für die neue LGSA-Filterfunktion zu erhalten, wurden die Weibull-Verteilungen der beteiligten Zufallsprozesse auf dem Weg in den LGSA-Bereich im Rahmen einer Fünf-Schritte-Herleitung so approximiert, dass am Ende die bedingte *a posteriori* Verteilungsdichtefunktion der LGSA-Koeffizienten berechnet werden konnte. Der Modalwert dieser Verteilung führte zum gesuchten recheneffizienten MAP-basierten LGSA-Schätzer. Um den verwendeten Approximationen und einigen Modellannahmen, die in der Realität verletzt werden, entgegenzuwirken, wurde anschließend eine Strategie zur Erhöhung der Flexibilität der statistischen Modellierung vorgeschlagen. Als Nebenprodukt entstand dabei eine modifizierte

Version des generalisierten MMSE-basierten PGSS-Schätzers aus [STCT98]. Anschließend folgte eine datengetriebene Parametrisierung der beiden Schätzer mit dem Ziel, die beste Qualität der entstörten Signale zu erreichen. Daraus resultierten zwei generalisierte spektrale Filterfunktionen: der modifizierte MMSE-PGSS-Schätzer und der vorgeschlagene MAP-LGSA-Schätzer.

Ein experimenteller Vergleich der beiden Filterfunktionen mit dem MMSE-LSA-Schätzer aus [EM85] offenbarte eine hervorragende Leistungsfähigkeit des MAP-basierten LGSA-Schätzers, der im Rahmen einer einkanaligen Sprachsignalentstörung der CHiME-3-Daten den MMSE-LSA-Schätzer hinsichtlich der Störsignaldämpfung im Mittel um etwa $1.4\,\mathrm{dB}$ (von $4.2\,\mathrm{dB}$ auf $5.6\,\mathrm{dB}$) übertraf und dabei keine nennenswerten Verluste in der Qualität der prozessierten Sprachsignale verursachte. Somit wurde gezeigt, dass der vorgeschlagene MAP-LGSA-Schätzer den Kompromiss zwischen der Sprachsignalqualität und der Störsignaldämpfung etwas besser löst als der MMSE-LSA- und der MMSE-PGSS-Schätzer.

9. Generalisiertes *Decision-Directed* Verfahren

Wie im letzten Kapitel vorgestellt, lässt sich die Leistungsfähigkeit der spektralen Sprachsignalentstörung durch den Einsatz von generalisierten modellbasierten Verfahren steigern, die im Bereich der generalisierten spektralen Amplituden entwickelt werden und nicht nur auf spektrale Amplituden oder Leistungen begrenzt sind. Nun soll davon auch der dritte Baustein eines Systems zur spektralen Sprachsignalentstörung aus Abb. 2.2 profitieren, welcher das *a priori* SNR schätzt. Wie in Abschnitt 2.2 beschrieben, spielt das *a priori* SNR als ein dominanter Parameter einer spektralen Filterfunktion die Rolle eines Dreh- und Angelpunktes in der spektralen Sprachsignalentstörung. Trotz Vorhandensein vieler ausgeklügelter Verfahren zur *a priori* SNR-Schätzung, von denen ein Bruchteil in Abschnitt 3.3 bereits vorgestellt wurde, genießt das *Decision-Directed* Verfahren aus [EM84] immer noch ein sehr hohes Ansehen. Außerdem dient das DD-Verfahren als Grundbaustein für Entwicklung seiner zahlreichen Modifikationen, die eine Milderung von bekannten Schwächen des DD-Verfahrens bezwecken. So wurden einige Versuche unternommen, den berüchtigten Nachhalleffekt des DD-Verfahrens zu mildern und somit seine Verfolgungseigenschaften zu verbessern.

Allerdings wird das *a priori* SNR $\xi(k, \ell)$, das mit Hilfe von spektralen Leistungen der beteiligten Zufallsprozesse in (2.22) definiert ist, von den konventionellen Verfahren meistens direkt im SNR-Bereich geschätzt. Dabei wird $\xi(k, \ell)$ bis auf einige wenigen Ausnahmen wie in [BGM08] oder in [EMTF15] aus dem *a posteriori* SNR $\gamma(k, \ell)$ berechnet, das mit (2.13) ebenfalls im SNR-Bereich definiert ist. Auch vom *Decision-Directed* Verfahren gibt es keine Modifikation, welche eine *a priori* SNR-Schätzung im Bereich der generalisierten spektralen Amplituden durchführt. Dabei gibt es keinen Nachweis darüber, ob der SNR-Bereich für eine *a priori* SNR-Schätzung optimal ist. Um diese Lücke zu füllen, wird in diesem Kapitel ein generalisiertes *Decision-Directed* (GDD) Verfahren entwickelt, das im generalisierten SNR-Bereich arbeitet, der im Weiteren über generalisierte spektrale Leistungen definiert wird. Das GDD-Verfahren wird adaptiv so gestaltet, dass es in Abhängigkeit vom eingangsseitigen globalen $\mathrm{SNR}_{\mathrm{IN}}$, den für die *a priori* SNR-Schätzung optimalen SNR-Bereich wählt.

Dafür werden zunächst in Abschnitt 9.1 generalisierte SNR-Größen samt ihrer statistischen Modellierung definiert. Angelehnt an das DD-Verfahren wird danach in Abschnitt 9.2 das generalisierte DD-Verfahren hergeleitet. Im darauf folgenden Abschnitt 9.3 wird das GDD-Verfahren im Rahmen einer datengetriebenen Parametrisierung adaptiv ausgelegt und zwar in den Experimenten mit den Sprachsignalen, die vom weißen Rauschen gestört werden. Um eine Überlegenheit des GDD-Verfahrens gegenüber dem DD-Verfahren beim Einsatz in einem System zur spektralen Sprachsignalentstörung zu zeigen, werden experimentelle Untersuchungen auf TIMIT- und CHiME-3-Daten in Abschnitt 9.4 durchgeführt, bevor dieses Kapitel mit einer kurzen Zusammenfassung in Abschnitt 9.5 schließt.

Man beachte, dass das generalisierte *Decision-Directed* Verfahren vom Autor dieser Arbeit zum ersten Mal in [CHU16] vorgestellt wurde.

9.1. Bereich der generalisierten SNR-Größen

Analog zur Definition des *a posteriori* SNR $\gamma(k,\ell)$ in (2.13) lässt sich ein generalisiertes *a posteriori* SNR einführen, der folgendermaßen definiert wird:

$$\gamma_\rho(k,\ell) \triangleq \frac{|Y(k,\ell)|^{2\rho}}{E[|D(k,\ell)|^{2\rho}]}, \tag{9.1}$$

wobei $\rho \in \mathbb{R}_{>0}$ ein reellwertiger positiver Exponent ist, mit dessen Hilfe sich der konventionelle SNR-Bereich auf den generalisierten SNR-Bereich erweitern lässt. Dabei ist das generalisierte *a posteriori* SNR $\gamma_\rho(k,\ell)$ genauso wie $\gamma(k,\ell)$ eine Zufallsvariable und geht für $\rho = 1$ in das konventionelle *a posteriori* SNR $\gamma(k,\ell)$ über. Man beachte, dass das generalisierte *a posteriori* SNR alternativ auch mit Hilfe des Kompressionsfaktors β aus (3.6) definiert werden könnte. Allerdings, da die SNR-Größen $\gamma(k,\ell)$ und $\xi(k,\ell)$ mit Hilfe der spektralen Leistungen (und nicht Amplituden) definiert werden, scheint die eingeführte Definition (9.1) mit dem Exponenten ρ für weitere Berechnungen sinnvoller zu sein.

Da bei den Herleitungen hier ähnlich wie in Abschnitt 2.1 angenommen wird, dass ein ungestörtes Sprachsignal $s(n)$ und eine Störung $d(n)$ additiv wie in (2.1) überlagert werden, gilt die Additivität der STFT-Koeffizienten (2.3) ebenfalls. Die Modellierung der STFT-Koeffizienten $S(k,\ell)$ und $D(k,\ell)$ mit (2.4) und (2.5) als zwei unkorrelierte komplexwertige mittelwertfreie normalverteilte Zufallsprozesse mit den frequenzabhängigen zeitvarianten Leistungsdichtespektren $\lambda_S(k,\ell)$ und $\lambda_D(k,\ell)$ aus (2.6) und (2.7) führt dazu, dass die STFT-Koeffizienten des gestörten Sprachsignals $Y(k,\ell)$ auch komplexwertig, mittelwertfrei und normalverteilt sind mit einer Verteilungsdichtefunktion (2.9). Unter diesen Annahmen unterliegen generalisierte spektrale Leistungen $|S(k,\ell)|^{2\rho}$ and $|D(k,\ell)|^{2\rho}$ ähnlich wie in Unterabschnitt 8.1.1 den reellwertigen Weibull-Verteilungen

$$p_{|S(k,\ell)|^{2\rho}}(s) = \text{Weib}(s;\, \lambda_S(k,\ell),\; 2\rho), \tag{9.2}$$
$$p_{|D(k,\ell)|^{2\rho}}(d) = \text{Weib}(d;\, \lambda_D(k,\ell),\; 2\rho). \tag{9.3}$$

Mit der Definition einer Weibull-Verteilung (8.5) hängen die Verteilungsdichtefunktionen (9.2) und (9.3) folgendermaßen vom Exponenten ρ ab:

$$\text{Weib}(x;\, \lambda_X, 2\rho) = \frac{1}{\rho\lambda_X} \cdot x^{\frac{1}{\rho}-1} \cdot e^{-\frac{1}{\lambda_X}\cdot x^{\frac{1}{\rho}}} \cdot U(x). \tag{9.4}$$

Die statistischen Momente κ-ter Ordnung werden mit (8.6) auch zu einer Funktion von ρ:

$$E[X^\kappa] = \lambda^{\kappa\cdot\rho} \cdot \Gamma(\kappa\cdot\rho + 1). \tag{9.5}$$

Man beachte, für $\rho = 1$ vereinfacht sich die Weibull-Verteilung (9.4) zu einer Exponentialverteilung wie in (5.1), die auch für statistische Modellierung des konventionellen *a posteriori* SNR $\gamma(k,\ell)$ wie in [Coh03] verwendet wird. Wie in Tab. 8.2 bereits erwähnt, ist die Weibull-Verteilung selbst ein Spezialfall einer generalisierten Gamma-Verteilung [KA10].

Unter Verwendung der Additivität (2.3) resultiert für die Verteilung der generalisierten spektralen Leistungen $|Y(k,\ell)|^{2\rho}$ des gestörten Sprachsignals auch eine Weibull-Verteilung:

$$p_{|Y(k,\ell)|^{2\rho}}(y) = \text{Weib}(y;\ \lambda_S(k,\ell) + \lambda_D(k,\ell),\ 2\rho). \tag{9.6}$$

Verwendet man (9.5) für (9.3) und führt eine Zufallsvariablentransformation von (9.1) mit (9.6) durch, resultiert für die Verteilungsdichtefunktion des generalisierten *a posteriori* SNR $\gamma_\rho(k,\ell)$ auch eine Weibull-Verteilung

$$p(\gamma_\rho) = \text{Weib}(\gamma_\rho;\ \lambda_{\gamma_\rho},\ 2\rho) \tag{9.7}$$

mit einem Skalierungsparameter

$$\lambda_{\gamma_\rho} = \frac{\xi(k,\ell)+1}{\sqrt[\rho]{\Gamma(\rho+1)}}. \tag{9.8}$$

Man beachte, dass während sich Skalierungsparameter von Weibull-Verteilungen der spektralen Amplituden (9.2) oder Leistungen (9.4) als ein Leistungsdichtespektrum darstellen, erweist sich λ_{γ_ρ} in (9.7) als eine skalierte SNR-Größe. Das zu schätzende *a priori* SNR $\xi(k,\ell)$ aus (2.22) fließt in (9.7) als ein Parameter ein und kann daher aus den Beobachtungen $\gamma_\rho(k,\ell)$ im Rahmen des GDD-Verfahrens geschätzt werden.

Verwendet man (9.5) für (9.3) in (9.1) resultiert für die Berechnung von $\gamma_\rho(k,\ell)$ aus dem konventionellen *a posteriori* SNR $\gamma(k,\ell)$ folgende Berechnungsvorschrift:

$$\gamma_\rho(k,\ell) = \frac{\gamma^\rho(k,\ell)}{\Gamma(\rho+1)}. \tag{9.9}$$

Analog zum generalisierten *a posteriori* SNR $\gamma_\rho(k,\ell)$ kann auch das generalisierte *a priori* SNR $\xi_\rho(k,\ell)$ definiert werden

$$\xi_\rho(k,\ell) \triangleq \frac{E[|S(k,\ell)|^{2\rho}]}{E[|D(k,\ell)|^{2\rho}]}, \tag{9.10}$$

das sich ähnlich der konventionellen Definition des *a priori* SNR (2.22) als Parameter und keine Zufallsvariable erweist. Verwendet man (9.5) für (9.2) und (9.3) in (9.10) kann $\xi_\rho(k,\ell)$ aus $\xi(k,\ell)$ wie folgt berechnet werden:

$$\xi_\rho(k,\ell) = \xi^\rho(k,\ell). \tag{9.11}$$

Bevor allerdings das GDD-Verfahren hergeleitet wird, soll eine alternative Definition des *a priori* SNR kurz diskutiert werden, das nicht als Parameter sondern als eine Zufallsvariable definiert wird. Einen Anlass dafür bietet das DD-Verfahren selbst, das von zwei Summanden lebt, dem ML-Schätzwert $\hat{\xi}^{\text{ML}}$ aus (3.12) und dem propagierten Schätzwert $\tilde{\xi}$ aus (3.14). Während ξ bei Herleitung von $\hat{\xi}^{\text{ML}}$ als (nichtzufälliger) Parameter aufgefasst wird, kann $\tilde{\xi}$ als Realisierung einer Zufallsvariablen angesehen werden. Eine weitere Motivation für eine Definition des *a priori* SNR als Zufallsvariable liefert die Publikation [MM80] – eine der ersten Veröffentlichungen, in welcher der Begriff des *a priori* SNR erwähnt wird. Denn in [MM80] wird das *a priori* SNR auch als Zufallsvariable definiert.

Also kann das *a priori* SNR im Gegensatz zu (2.22) und ähnlich wie (3.14) auch als eine Zufallsvariable definiert werden und zwar:

$$\zeta(k,\ell) \triangleq \frac{|S(k,\ell)|^2}{\lambda_D(k,\ell)}, \tag{9.12}$$

das im Rahmen dieser Arbeit als ein momentanes *a priori* SNR bezeichnet wird. Die beiden Definitionen des *a priori* SNR aus (2.22) und (9.12) sind miteinander ganz einfach verknüpft und zwar über $\xi(k,\ell) = E\left[\zeta(k,\ell)\right]$. Ähnlich dem konventionellen DD-Verfahren soll auch das generalisierte *Decision-Directed* Verfahren von beiden Definitionen des *a priori* SNR $\xi(k,\ell)$ in (2.22) und in (9.12) Gebrauch machen. Dafür ist allerdings die Generalisierung von $\zeta(k,\ell)$ nötig, die ja im konventionellen SNR-Bereich definiert ist. Das Gegenstück von $\zeta(k,\ell)$ im generalisierten SNR-Bereich soll ein generalisiertes momentanes *a priori* SNR sein, das folgendermaßen definiert wird:

$$\zeta_\rho(k,\ell) \triangleq \frac{|S(k,\ell)|^{2\rho}}{E[|D(k,\ell)|^{2\rho}]} = \frac{|S(k,\ell)|^{2\rho}}{\lambda_D^\rho(k,\ell)\cdot\Gamma(\rho+1)}. \tag{9.13}$$

Analog wie $\gamma_\rho(k,\ell)$ aus $\gamma(k,\ell)$ mit (9.9) berechnet wird, kann $\zeta_\rho(k,\ell)$ aus $\zeta(k,\ell)$ berechnet werden und zwar mit:

$$\zeta_\rho(k,\ell) = \frac{\zeta^\rho(k,\ell)}{\Gamma(\rho+1)}. \tag{9.14}$$

Aus (9.2) und (9.3) resultiert mit (9.5) für die Verteilungsdichtefunktion von $\zeta_\rho(k,\ell)$ wieder eine Weibull-Verteilung

$$p(\zeta_\rho) = \text{Weib}(\zeta_\rho;\ \lambda_{\zeta_\rho},\ 2\rho), \tag{9.15}$$

mit einem Skalierungsparameter

$$\lambda_{\zeta_\rho} = \frac{\xi(k,\ell)}{\sqrt[\rho]{\Gamma(\rho+1)}}, \tag{9.16}$$

der das zu schätzende *a priori* SNR $\xi(k,\ell)$ als Parameter beinhaltet. Da allerdings weder $\zeta(k,\ell)$ aus (9.12) noch ζ_ρ aus (9.13) in einer Anwendung direkt beobachtbar sind, lässt sich $\xi(k,\ell)$ nicht ohne Weiteres aus den beiden Größen schätzen. Das generalisierte *a priori* SNR aus (9.10) als Parameter ist mit dem generalisierten momentanen *a priori* SNR aus (9.13) als Zufallsvariable über $\xi_\rho(k,\ell) = E\left[\zeta_\rho(k,\ell)\right]$ verknüpft.

Nachdem alle notwendigen Definitionen eingeführt wurden, kann das GDD-Verfahren vorgestellt werden.

9.2. Das generalisierte *Decision-Directed* Verfahren

Beim Herleiten des GDD-Verfahrens im generalisierten SNR-Bereich wird von der gewichteten Summe (3.13) des DD-Verfahrens ausgegangen, mit dem einzigen Unterschied, dass sie im Bereich der generalisierten SNR-Größen aufgestellt wird:

$$\hat{\theta}_\rho^{\text{DD}}(k,\ell) = \alpha_\rho \cdot \tilde{\zeta}_\rho(k,\ell-1) + (1-\alpha_\rho)\cdot\hat{\xi}_\rho^{\text{ML}}(k,\ell), \tag{9.17}$$

wobei $\hat{\theta}_{\rho}^{\mathrm{DD}}(k,\ell)$ ein DD-Schätzwert des generalisierten *a priori* SNR und $\alpha_{\rho} \in \mathbb{R}_{[0;1]}$ ein Gewichtsfaktor des GDD-Verfahrens sind. Da die Größe $\hat{\theta}_{\rho}^{\mathrm{DD}}(k,\ell)$ im Rahmen des GDD-Verfahrens entweder als ein Schätzwert eines Parameters oder als eine Realisierung einer Zufallsvariable aufgefasst wird, wird sie mit Hilfe eines neuen Symbols θ bezeichnet und nicht mit ζ oder ξ. Außerdem beinhaltet der erste Summand in (9.17) einen generalisierten propagierten *a priori* SNR-Schätzwert

$$\tilde{\zeta}_{\rho}(k,\ell-1) = \frac{|\hat{S}(k,\ell-1)|^{2\rho}}{\lambda_D^{\rho}(k,\ell-1)\cdot\Gamma(\rho+1)} = G_{H_1}^{2\rho}(k,\ell-1)\cdot\hat{\gamma}_{\rho}(k,\ell-1), \tag{9.18}$$

der als eine Realisierung des generalisierten momentanen *a priori* SNR $\zeta_{\rho}(k,\ell)$ aus (9.13) aufgefasst werden kann. Und da $\zeta_{\rho}(k,\ell)$ als eine Zufallsvariable definiert ist, wird das generalisierte propagierte *a priori* SNR hier mit dem Symbol ζ bezeichnet und nicht mit dem Symbol ξ, wie dies beim DD-Verfahren aus (3.13) der Fall ist. Alternativ kann $\tilde{\zeta}_{\rho}(k,\ell-1)$ über $\tilde{\xi}(k,\ell-1)$ aus (3.14) wie folgt berechnet werden:

$$\tilde{\zeta}_{\rho}(k,\ell-1) = \frac{\tilde{\xi}^{\rho}(k,\ell-1)}{\Gamma(\rho+1)}. \tag{9.19}$$

Sonst enthält der zweite Summand in (9.17) einen *Maximum-Likelihood*-Schätzwert $\hat{\xi}_{\rho}^{\mathrm{ML}}(k,\ell)$ des generalisierten *a priori* SNR $\xi_{\rho}(k,\ell)$ aus (9.10), der über (9.11) und (9.8) in die Weibull-Verteilung (9.7) einfließt. Als Schätzwert eines Parameters lässt sich $\hat{\xi}_{\rho}^{\mathrm{ML}}(k,\ell)$ aus einer Beobachtung $\hat{\gamma}_{\rho}(k,\ell)$ berechnen

$$\hat{\xi}_{\rho}^{\mathrm{ML}}(k,\ell) = \left(\max\left(\sqrt[\rho]{\Gamma(\rho+1)\cdot\hat{\gamma}_{\rho}(k,\ell)} - 1\,,\, 0 \right) \right)^{\rho}. \tag{9.20}$$

Mit (9.9) kann $\hat{\xi}_{\rho}^{\mathrm{ML}}(k,\ell)$ auch direkt aus einem *a posteriori* SNR $\gamma(k,\ell)$ berechnet werden

$$\hat{\xi}_{\rho}^{\mathrm{ML}}(k,\ell) = \left(\max\left(\hat{\gamma}(k,\ell) - 1\,,\, 0 \right) \right)^{\rho}. \tag{9.21}$$

Wie man der Gleichung (9.17) entnehmen kann, wird im Rahmen des GDD-Verfahrens nichts anderes als das DD-Verfahren im generalisierten SNR-Bereich implementiert, in dem die Gewichtung der Terme $\tilde{\zeta}_{\rho}(k,\ell-1)$ und $\hat{\xi}_{\rho}^{\mathrm{ML}}(k,\ell)$ stattfindet. Und da (9.17), wie auch (3.13) in Abschnitt 3.3, mit einer rekursiven Glättung erster Ordnung verglichen werden kann, lässt sich festhalten, dass das GDD-Verfahren im Unterschied zum DD-Verfahren die rekursive Filterung nicht mehr im konventionellen SNR-Bereich sondern im generalisierten SNR-Bereich durchführt. Die Vorteile dieser Filterung werden in den experimentellen Untersuchungen ersichtlich, die im Weiteren vorgestellt werden. Zunächst allerdings soll erläutert werden, wie das eigentlich gesuchte *a priori* SNR $\hat{\xi}^{\mathrm{GDD}}(k,\ell)$ des GDD-Verfahrens aus $\hat{\theta}_{\rho}^{\mathrm{DD}}(k,\ell)$ aus (9.17) berechnet wird. Dafür ist es ausschlaggebend, welche Rolle der Größe $\theta_{\rho}^{\mathrm{DD}}(k,\ell)$ zugeschrieben wird, die eines Parameters wie in (9.10) oder die einer Zufallsvariablen wie in (9.12).

Wird $\theta_{\rho}^{\mathrm{DD}}(k,\ell)$ als ein Parameter betrachtet, lässt sich das resultierende *a priori* SNR des GDD-Verfahrens $\hat{\xi}^{\mathrm{GDD}}(k,\ell)$ unter Verwendung von (9.11) wie folgt berechnet werden:

$$\hat{\xi}^{\mathrm{GDD}}(k,\ell) = \max\left(\left(\hat{\theta}_{\rho}^{\mathrm{DD}}(k,\ell) \right)^{\frac{1}{\rho}},\ \xi_{\min} \right). \tag{9.22}$$

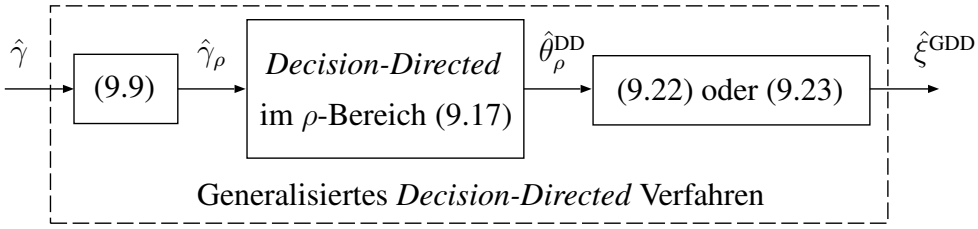


Abbildung 9.1.: Blockschaltbild des generalisierten *Decision-Directed* Verfahrens.

ξ_{min} begrenzt dabei den *a priori* SNR-Schätzwert von unten ähnlich wie in (2.25). Betrachtet man $\theta_\rho^{\mathrm{DD}}(k,\ell)$ als eine Zufallsvariable kann für die Berechnung von $\hat{\xi}^{\mathrm{GDD}}(k,\ell)$ die Gleichung (9.14) herangezogen werden, die in folgender Berechnungsvorschrift resultiert

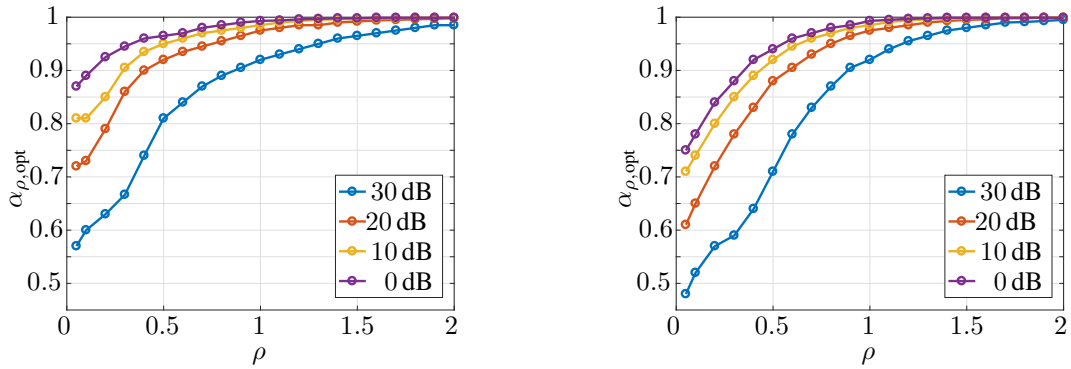
$$\hat{\xi}^{\mathrm{GDD}}(k,\ell) = \max\left(\left(\hat{\theta}_\rho^{\mathrm{DD}}(k,\ell)\cdot\Gamma(\rho+1)\right)^{\frac{1}{\rho}},\ \xi_{\mathrm{min}}\right). \tag{9.23}$$

Aufschluss darüber, welche der beiden Gleichungen (9.22) oder (9.23) im Rahmen des GDD-Verfahrens verwendet werden soll, sollen bevorstehende experimentelle Untersuchungen bringen, die in Abschnitt 9.3 vorgestellt werden.

Zusammenfassend lässt sich das vorgestellte GDD-Verfahren in drei separate Berechnungsschritte aufteilen, die in Abb. 9.1 dargestellt sind. Im ersten Schritt geht man mit Hilfe von (9.9) aus dem konventionellen SNR-Bereich in den Bereich der generalisierten SNR-Größen. Hier wird im Rahmen des zweiten Schrittes das konventionelle DD-Verfahren realisiert und $\hat{\theta}_\rho^{\mathrm{DD}}$ aus $\hat{\gamma}_\rho$ mit (9.17) berechnet. Im dritten Schritt bestimmt man den endgültigen *a priori* SNR-Schätzwert $\hat{\xi}^{\mathrm{GDD}}$, wofür entweder (9.22) oder (9.23) verwendet werden. Man beachte, dass das vorgestellte GDD-Verfahren für $\rho = 1$ zum konventionellen DD-Verfahren wird, unabhängig davon, ob man (9.22) oder (9.23) verwendet. Wie experimentelle Untersuchungen zeigten, ist $\hat{\zeta}_\rho^{\mathrm{GDD}}(k,1) = \max((\hat{\gamma}(k,1)-1)^\rho,\ 0)$ für eine Initialisierung gut geeignet. Während das GDD-Verfahren dieselbe untere Grenze ξ_{min} wie auch das DD-Verfahren verwenden kann, müssen sein Exponent ρ und sein Gewichtsfaktor α_ρ noch angemessen gewählt werden. Für eine gute Wahl dieser Parameter werden Experimente mit Sprachsignalen durchgeführt, die von einem additiven weißen Störsignal überlagert werden.

9.3. Datengetriebene Parametrisierung des GDD-Verfahrens

Nun sollen die Parameter des GDD-Verfahrens ρ und α_ρ so gewählt werden, dass die entstörten Sprachsignale die bestmögliche Sprachsignalqualität erreichen, wenn das GDD-Verfahren in einem System zur spektralen Sprachsignalentstörung eingesetzt wird. Dafür soll eine Experimentalreihe mit den Sprachsignalen der TIMIT-Datenbank und mit dem weißen Rauschen der SPIB-Datenbank durchgeführt werden, die in Abschnitt 2.4 bereits beschrieben wurden. Vor einer Signalverarbeitung werden die kurzen Sprachsignale der TIMIT-Datenbank mit weiblichen und männlichen Sprechern jeweils zu den längeren Signalen der Länge von einer Minute zusammengesetzt, bevor sie mit dem weißen Rauschen bei vier verschiedenen Werten des globalen eingangsseitigen $\mathrm{SNR}_{\mathrm{IN}} \in \{0, 10, 20, 30\,\mathrm{dB}\}$ überlagert werden,



(a) Verwendung von (9.22) (b) Verwendung von (9.23)

Abbildung 9.2.: Optimale Gewichtsfaktoren $\alpha_{\rho,\mathrm{opt}}$ für die beste Qualität der entstörten Sprachsignale gemessen in MOS-LQO$_{\mathrm{WB}}$ als Funktion des Exponenten ρ für die Werte von $\mathrm{SNR}_{\mathrm{IN}} \in \{0, 10, 20, 30\,\mathrm{dB}\}$: (a) Verwendung von (9.22) und (b) Verwendung von (9.23).

das im Weiteren mit Υ gekennzeichnet wird. Alle Signale weisen die Abtastrate von 16 kHz auf, sodass für die Messung der Sprachsignalqualität das breitbandige MOS-LQO$_{\mathrm{WB}}$-Maß verwendet werden kann, das in Abschnitt 2.3 bereits eingeführt wurde. Für eine STFT aus Abschnitt 2.1 werden ein Hanning-Analysefenster der Länge $K = 512$ Abtastwerte und ein Rahmenvorschub von $R = 128$ Abtastwerte verwendet. Die Schätzung der Rauschleistungsdichte $\lambda_D(k, \ell)$ wird mit dem *Minimum Statistics* Verfahren aus [Mar01] realisiert, wobei die Länge des Fensters für eine effiziente Minimumsuche zu $D_{\mathrm{MS}} = 96$ STFT-Rahmen gewählt, die in $U_{\mathrm{MS}} = 8$ Unterfenster der Länge von $V_{\mathrm{MS}} = 12$ Rahmen aufgeteilt wird. Als spektrale Filterfunktion wird der MMSE-LSA-Schätzer aus [EM85] mit $G_{\mathrm{min}} = -25\,\mathrm{dB}$ verwendet [Coh01]. Auf den vierten Baustein des Systems zur spektralen Sprachsignalentstörung aus Abb. 2.2 wird in den Experimenten dieses Kapitels verzichtet. Für eine *a priori* SNR-Schätzung wird das in Abschnitt 9.2 vorgeschlagene GDD-Verfahren eingesetzt, das für $\rho = 1$ zum DD-Verfahren wird. Dabei wird der minimale *a priori* SNR-Schätzwert ξ_{min} zu $-25\,\mathrm{dB}$ gesetzt [Coh04]. Die GDD-Parameter ρ und α_ρ werden im Rahmen der Experimente jeweils in den Wertebereichen $[0.05; 2]$ und $[0; 1]$ variiert.

Die durchgeführten Untersuchungen zeigten, dass das GDD-Verfahren zu den entstörten Sprachsignalen mit bester Qualität gemessen in MOS-LQO$_{\mathrm{WB}}$-Werten führt, wenn der Gewichtsfaktor α_ρ in Abhängigkeit vom verwendeten Exponenten ρ und vom vorliegenden $\mathrm{SNR}_{\mathrm{IN}}$-Wert Υ gewählt wird. Die optimalen Gewichtsfaktoren $\alpha_{\rho,\mathrm{opt}}$ sind als Funktion von ρ für verschiedene $\mathrm{SNR}_{\mathrm{IN}}$-Werte in Abb. 9.2 dargestellt. Die resultierenden $\alpha_{\rho,\mathrm{opt}}$ Werte für $\rho = 1$ sind in Tab. 9.1 angegeben. Wie man sieht, soll der optimale Gewichtsfaktor mit steigenden $\mathrm{SNR}_{\mathrm{IN}}$-Werten eigentlich kleiner gewählt werden. Oder mit anderen Worten, je weniger die Sprachsignale gestört sind, desto höher darf das Gewicht des ML-Schätzers in (3.13) gewählt werden. Allerdings wird im konventionellen DD-Verfahren nach [EM84] von einem adaptiven Gewichtsfaktor kein Gebrauch gemacht. Stattdessen wird häufig $\alpha_{\mathrm{DD}} = 0.98$ ge-

$\mathrm{SNR}_{\mathrm{IN}}$ / dB	0	10	20	30
$\alpha_{\mathrm{DD,opt}}$	0.993	0.985	0.975	0.92

Tabelle 9.1.: Optimale Gewichtsfaktoren des DD-Verfahrens für beste Sprachsignalqualität.

wählt, der laut der angestellten Untersuchung für $\mathrm{SNR}_{\mathrm{IN}}$ von etwa 15 dB optimal ist. Auch für die Werte $\rho \neq 1$ behält $\alpha_{\rho,\mathrm{opt}}$ seinen fallenden Charakter als Funktion von $\mathrm{SNR}_{\mathrm{IN}}$, wie man Abb. 9.2 entnehmen kann. Anders ist es, wenn man $\alpha_{\rho,\mathrm{opt}}$ als Funktion von ρ für verschiedene feste $\mathrm{SNR}_{\mathrm{IN}}$-Werte betrachtet, denn hier steigt $\alpha_{\rho,\mathrm{opt}}$ mit größeren Werten von ρ und zwar unabhängig davon, ob man im letzten Berechnungsschritt (9.22) oder (9.23) verwendet. Im generalisierten SNR-Bereich scheint es also für $\rho < 1$ möglich zu sein, den Gewichtsfaktor α_ρ abzusenken, was eine Milderung des Nachhalleffektes mit sich bringen kann. Außerdem ist zu erwähnen, dass die Verwendung von (9.23) generell kleinere Werte von $\alpha_{\rho,\mathrm{opt}}$ erfordert als Benutzung von (9.22), wie man in Abb. 9.2 sieht.

Um zu verdeutlichen, welche Verbesserung in der Sprachsignalqualität das GDD-Verfahren gegenüber dem DD-Verfahren prinzipiell erreichen kann, wird eine differentielle Messgröße für die Sprachsignalqualität definiert

$$\Delta\text{MOS-LQO}_{\mathrm{GDD}} = \text{MOS-LQO}_{\mathrm{WB,GDD}} - \text{MOS-LQO}_{\mathrm{WB,DD}}. \tag{9.24}$$

In den Experimenten wird $\Delta\text{MOS-LQO}_{\mathrm{GDD}}$ für den Fall berechnet, dass ein optimaler Gewichtsfaktor $\alpha_{\rho,\mathrm{opt}}$ im GDD-Verfahren verwendet wird, dessen Abhängigkeit von ρ und Υ in Abb. 9.2 vorgestellt wurde. Die Gewichtsfaktoren des DD-Verfahrens werden dabei entsprechend den simulierten $\mathrm{SNR}_{\mathrm{IN}}$-Werten aus Tab. 9.1 auch optimal gesetzt. Die resultierenden $\Delta\text{MOS-LQO}_{\mathrm{GDD}}$-Werte sind in Abb. 9.3 als Funktion vom Exponenten ρ für verschiedene $\mathrm{SNR}_{\mathrm{IN}}$-Werte dargestellt, wobei in Abb. 9.3 (a) die Gleichung (9.22) und in Abb. 9.3 (b) die Gleichung (9.23) verwendet wurde. Dabei fällt auf, dass die beiden Versionen des GDD-Verfahrens eine etwas unterschiedliche Verbesserung der Sprachsignalqualität erreichen. Ent-

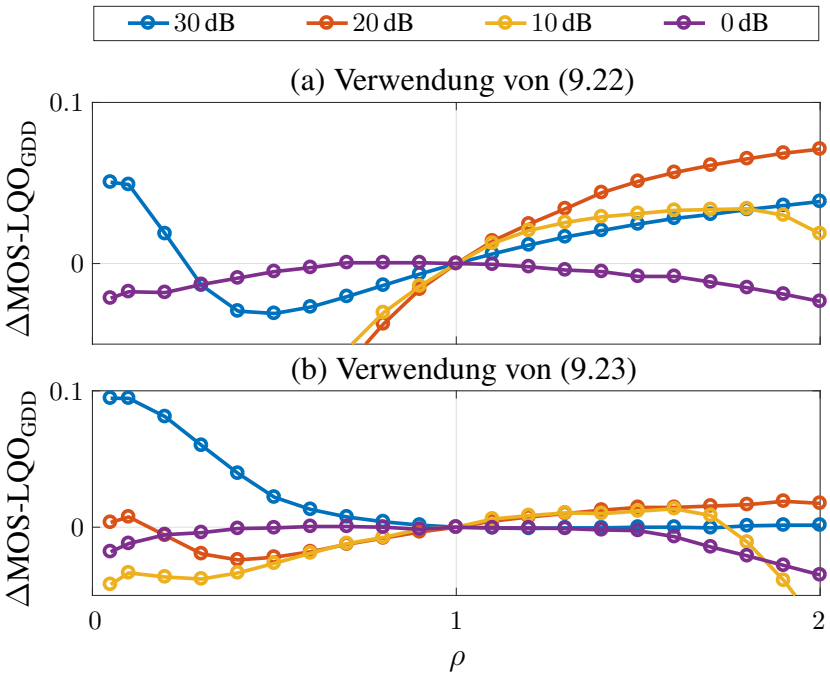


Abbildung 9.3.: Verbesserung der Sprachsignalqualität gemessen in $\Delta\text{MOS-LQO}_{\mathrm{GDD}}$ Werten beim Einsatz des GDD-Verfahrens im Vergleich zum DD-Verfahren als Funktion des Exponenten $\rho \in [0.05; 2]$ in Experimenten mit Sprachsignalen gestört von weißem Störsignal bei $\mathrm{SNR}_{\mathrm{IN}} \in \{0, 10, 20, 30\,\mathrm{dB}\}$: (a) Verwendung von (9.22), (b) Verwendung von (9.23).

sprechend Abb. 9.3 (a) führt das GDD-Verfahren mit (9.22) zu einer leichten Verbesserung der Sprachsignalqualität bei den positiven SNR_{IN}-Werten für $\rho > 1$. Während die $\Delta\text{MOS-LQO}_{\text{GDD}}$-Werte hier am größten für $20\,\text{dB}$ ausfallen, kann das DD-Verfahren für stark gestörte Sprachsignale wie bei $\text{SNR}_{\text{IN}} = 0\,\text{dB}$ vom GDD-Verfahren nicht übertroffen werden, denn alle positiven $\Delta\text{MOS-LQO}_{\text{GDD}}$-Werte sind in diesem Fall nicht nennenswert. Bei einer Realisierung des GDD-Verfahrens mit (9.23) wird eine Verbesserung der Sprachsignalqualität im Wesentlichen nur für die wenig gestörten Sprachsignale beobachtet, wenn verhältnismäßig kleine Werte des Exponenten um $\rho \approx 0.1$ verwendet werden. Ein Grund dafür sind relativ glatte Verläufe der *a priori* SNR-Schätzwerte in Sprachsignalabwesenheit, denn das GDD-Verfahren arbeitet in diesem Fall auf den sehr stark komprimierten generalisierten SNR-Größen. Gleichzeitig besitzt das GDD-Verfahren aufgrund der kleinen Gewichtsfaktoren eine hohe Verfolgungsfähigkeit und ist im Stande, steigenden und fallenden LDS-Flanken des ungestörten Sprachsignals schnell nachzukommen. Ähnliche Eigenschaften des GDD-Verfahrens für kleine Werte von ρ sind auch in Abb. 9.3 (a) für $\text{SNR}_{\text{IN}} = 30\,\text{dB}$ zu sehen, allerdings fallen dort $\Delta\text{MOS-LQO}_{\text{GDD}}$-Werte kleiner als in Abb. 9.3 (b) aus. Außerdem bestätigen die beiden Auswertungen in Abb. 9.3 die Tatsache, dass für die sehr stark gestörten Sprachsignale $\rho = 1$ wirklich die beste Wahl hinsichtlich der Sprachsignalqualität ist. Somit wird experimentell gezeigt, dass das DD-Verfahren bei stark gestörten Sprachsignalen bereits eine gute Leistungsfähigkeit zeigt.

Die durchgeführte Untersuchung des GDD-Verfahrens zeigte auch, dass es keinen universell optimalen Exponenten gibt, der unabhängig vom vorliegenden SNR_{IN}-Wert die Qualität der prozessierten Sprachsignale immer nur verbessert. Aus diesem Grund scheint es eine gute Lösung für die Parametrisierung des GDD-Verfahrens zu sein, die Parameter ρ und α_ρ in Abhängigkeit vom globalen eingangsseitigen SNR_{IN} zu wählen, der in diesem Fall zusätzlich geschätzt werden muss. In einer praktischen Realisierung der adaptierten Parameter $\rho(\Upsilon)$ und $\alpha_\rho(\Upsilon)$ kann dabei SNR_{IN} aus den resultierenden *a priori* SNR-Schätzwerten folgendermaßen geschätzt werden:

$$\hat{\Upsilon}_0(\ell) = \frac{\ell - 1}{\ell} \cdot \hat{\Upsilon}_0(\ell - 1) + \frac{1}{\ell} \cdot \frac{1}{K/2 + 1} \sum_{k=0}^{K/2} \hat{\xi}^{\text{GDD}}(k, \ell), \tag{9.25}$$

wobei $\Upsilon_0 = 10^{\Upsilon/10}$ und K jeweils der absolute SNR_{IN}-Wert und die FFT-Länge aus (2.2) sind. Für die Initialisierung wird $\hat{\Upsilon}(1) = 15\,\text{dB}$ umgerechnet in einen absoluten Wert $\hat{\Upsilon}_0(1)$ verwendet. Man beachte, dass die Verwendung von adaptiven Parametern, die über die Schätzwerte eines globalen eingangsseitigen SNR-Wertes gesteuert werden, eine gängige Technik

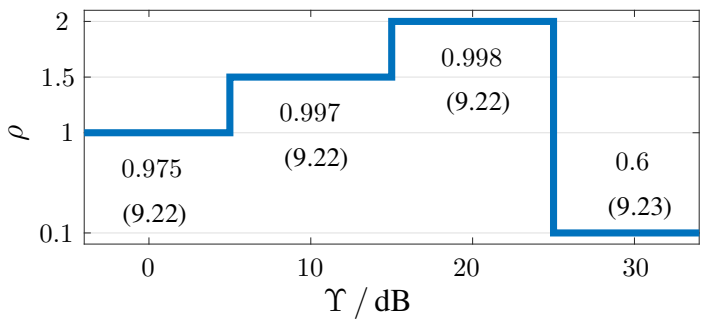


Abbildung 9.4.: Adaptionsfunktion $\rho(\Upsilon)$ des GDD-Verfahrens.

in der spektralen Sprachsignalentstörung ist. Einige für die vorgestellte Problemstellung geeignete Beispiele dafür sind [MMCA04, YKR05, LSH$^+$08]. Zieht man experimentelle Ergebnisse aus Abb. 9.2 und Abb. 9.3 in Betracht, kann eine treppenförmige Funktion $\rho(\Upsilon)$ aus Abb. 9.4 als eine Adaptionsfunktion verwendet werden. Die Werte des empfohlenen Gewichtsfaktors $\alpha_\rho(\Upsilon)$ sind auch in Abb. 9.4 angegeben. Man beachte, dass das adaptive GDD-Verfahren sowohl von (9.22) als auch von (9.23) Gebrauch macht, um einen maximalen Gewinn in der Sprachsignalqualität zu erzielen. Prinzipiell könnten auch monotone Funktionen $\rho(\Upsilon)$ und $\alpha_\rho(\Upsilon)$ als Adaptionsfunktionen definiert werden. Allerdings wird in weiteren Experimenten darauf verzichtet und die treppenförmige Funktion aus Abb. 9.4 verwendet.

9.4. Experimente auf TIMIT/SPIB- und CHiME-3- Daten

Nachdem das GDD-Verfahren mit den adaptiven Parametern ausgestattet wurde, soll seine Leistungsfähigkeit im Vergleich zum DD-Verfahren auf größeren Datensätzen getestet werden. Dafür werden zwei Untersuchungen angestellt, eine auf den TIMIT-Daten überlagert mit den Störsignalen verschiedener Rauschtypen der SPIB-Datenbank und die andere auf den CHiME-3-Daten, die in Abschnitt 2.4 beschrieben wurden. In den beiden Experimentalreihen kommt dasselbe System zur spektralen Sprachsignalentstörung zum Einsatz, das auch für die Parametrisierung des GDD-Verfahrens in Abschnitt 9.3 eingesetzt wurde, mit dem einzigen Unterschied, dass der Gewichtsfaktor des DD-Verfahrens hier nicht entsprechend Tab. 9.1 sondern ähnlich wie in [EM85] konstant zu $\alpha_{\text{DD}} = 0.975$ gewählt wird. Da eine Verbesserung der Qualität der prozessierten Sprachsignale gleichzeitig eine Reduktion einer Störsignaldämpfung verursachen kann, wie z. B. in Abb. 7.17 ersichtlich ist, wird neben dem ΔMOS-LQO$_{\text{GDD}}$-Maß aus (9.26) auch ein differentielles Maß für eine Veränderung einer Störsignaldämpfung definiert

$$\Delta\text{SNR}_{\text{GDD}} = \text{SNR}_{\text{OUT,GDD}} - \text{SNR}_{\text{OUT,DD}}, \tag{9.26}$$

wobei SNR$_{\text{OUT,GDD}}$ und SNR$_{\text{OUT,DD}}$ globale ausgangsseitige SNR-Werte aus (2.41) sind, wenn entweder das GDD-Verfahren oder das konventionelle DD-Verfahren als ein *a priori* SNR-Schätzer eingesetzt werden. Betrachtet man die beiden Messgrößen ΔMOS-LQO$_{\text{GDD}}$ und ΔSNR$_{\text{GDD}}$ zusammen, lässt sich das Leistungsvermögen des vorgeschlagenen GDD-Verfahrens im Vergleich zum DD-Verfahren besser beurteilen.

Gestörte Sprachsignale der ersten Untersuchung werden ähnlich wie die Daten erzeugt, welche für die Parametrisierung des GDD-Verfahrens verwendet wurden, allerdings mit einem einzigen Unterschied, dass neben dem weißen Rauschen noch 14 weitere Rauschtypen

SNR$_{\text{IN}}$ / dB	0	10	20	30	AVG
ΔMOS-LQO$_{\text{GDD}}$	0	0.02	0.05	0.15	0.06
ΔSNR$_{\text{GDD}}$ / dB	0.61	1.03	1.19	2.89	1.43

Tabelle 9.2.: Mittlere Verbesserung der Sprachsignalqualität und der Störsignaldämpfung in den Untersuchungen auf den TIMIT-Sprachsignalen überlagert mit den Störsignalen der SPIB-Datenbank für verschiedene SNR$_{\text{IN}}$-Werte.

der SPIB-Datenbank als Störung eingesetzt werden. Die resultierenden ΔMOS-LQO$_{\text{GDD}}$- und ΔSNR$_{\text{GDD}}$-Werte, welche zum Einen über Ergebnisse auf den Signalen mit den weiblichen und männlichen Sprechern und zum Anderen über Ergebnisse für alle verwendeten Rauschtypen gemittelt werden, sind in Tab. 9.2 für verschiedene SNR$_{\text{IN}}$-Werte angegeben. Wie man sieht, wird das DD-Verfahren für SNR$_{\text{IN}} = 0\,$dB vom GDD-Verfahren hinsichtlich der Sprachsignalqualität nicht übertroffen, was aus den Experimenten in Abschnitt 9.3 bereits bekannt ist. Allerdings trägt der vorgeschlagene *a priori* SNR-Schätzer hier zur um etwa $0.6\,$dB besseren Störsignalunterdrückung bei, welche sicherlich auf den adaptiven Charakter seiner Parameter zurückzuführen ist. Werden Sprachsignale der TIMIT-Datenbank weniger gestört, sorgt das GDD-Verfahren für die prozessierten Signale mit leicht besserer Qualität, als dies beim DD-Verfahren der Fall ist. Dabei ist die Qualitätsverbesserung um so größer, je weniger Störung im gestörten Signal vorliegt. Daraus kann man schließen, dass das vorgeschlagene GDD-Verfahren die leicht gestörten Sprachsignale weniger verzerrt oder angreift als das DD-Verfahren. Erstaunlicherweise kommt die erzielte Verbesserung der Sprachsignalqualität nicht auf Kosten einer schwächeren Störsignaldämpfung zustande. Ganz im Gegenteil, mit Verbesserung der Signalqualität wächst auch der Gewinn in der Rauschunterdrückung. Während die Qualitätsverbesserung einen durchschnittlichen Wert von etwa 0.06 Punkten auf der MOS-Skala erreicht, beläuft sich die mittlere Erhöhung der Störsignaldämpfung auf etwa $1.43\,$dB. Die letztere bekommt einen noch höheren Stellenwert, wenn man beachtet, dass das GDD-Verfahren einen durchschnittlichen ΔSNR-Wert des DD-Verfahrens von $4.43\,$dB, der entsprechend (2.42) definiert ist, bis auf $5.86\,$dB verbessert.

Betrachtet man die Leistungsfähigkeit des GDD-Verfahrens für unterschiedliche Rauschtypen, ergeben sich ΔMOS-LQO$_{\text{GDD}}$- und ΔSNR$_{\text{GDD}}$-Werte, die in Abb. 9.5 dargestellt sind. Die resultierenden Messgrößen ΔMOS-LQO$_{\text{GDD}}$ und ΔSNR$_{\text{GDD}}$ werden für die Darstellung hier sowohl über die Ergebnisse der weiblichen und männlichen Sprechern als auch über die der simulierten SNR$_{\text{IN}}$-Werte $\{0, 10, 20, 30\ \text{dB}\}$ gemittelt. Die erzielten Bewertungsmaße zeigen, dass der vorgeschlagene *a priori* SNR-Schätzer besonders hinsichtlich der Störsi-

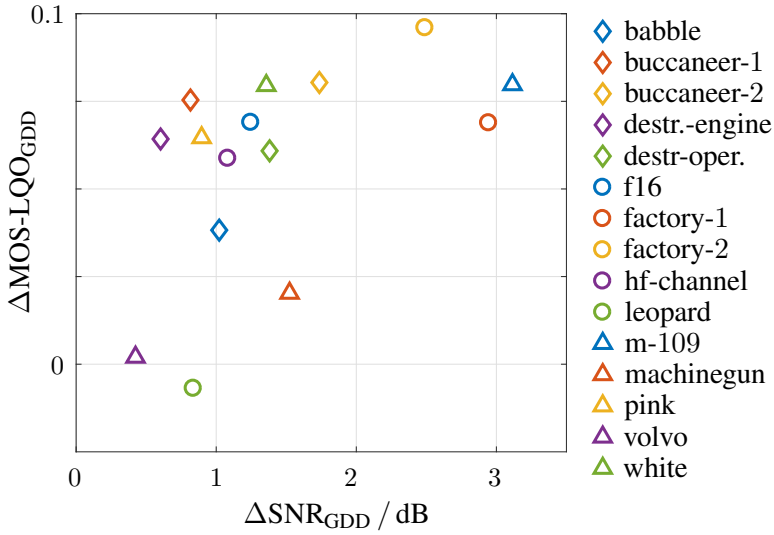


Abbildung 9.5.: Qualitätsverbesserung in ΔMOS-LQO$_{\text{GDD}}$ gegenüber Störsignaldämpfung in ΔSNR$_{\text{GDD}}$ für TIMIT-Sprachsignale gestört von verschiedenen Störsignalen der SPIB-Datenbank gemittelt über die globalen eingangsseitigen SNR$_{\text{IN}}$-Werte $\{0, 10, 20, 30\ \text{dB}\}$.

-		*bus*	*caf*	*ped*	*str*	*avg*
ΔMOS-LQO$_{\text{GDD}}$		0.01	0	0.01	0.02	0.01
ΔSNR$_{\text{GDD}}$ / dB		1.01	0.75	0.77	0.94	0.86
SNR$_{\text{OUT}}$ / dB	DD	7.37	10.24	12.14	8.80	9.64
	GDD	8.36	10.99	12.91	9.74	10.50

Tabelle 9.3.: Verbesserung der Sprachsignalqualität und der Störsignaldämpfung für verschiedene Störumgebungen in den Untersuchungen auf CHiME-3-Daten.

gnaldämpfung vorteilhaft ist, denn alle ΔSNR$_{\text{GDD}}$-Werte fallen positiv aus. Aber auch in der Sprachsignalqualität übertrifft das GDD-Verfahren das konventionelle DD-Verfahren, außer bei den tiefpasslastigen Rauschtypen wie '*leopard*' und '*volvo*'. Besonders erwähnenswert sind hier die guten Bewertungsmaße nichtstationärer Rauschtypen wie '*m-109*', '*factory-1*' und '*factory-2*'. Insgesamt zeigt die durchgeführte Untersuchung auf TIMIT- und SPIB-Daten, dass die Überlegenheit des GDD-Verfahrens gegenüber dem konventionellen DD-Verfahren nicht nur beim weißen Rauschen vorhanden ist, sondern sich auch bei nichtstationären Rauschtypen beobachten lässt. Um herauszufinden, ob sie auch bei den Sprachsignalen vorhanden ist, die an der Parametrisierung des GDD-Verfahrens nicht teilnahmen, soll eine weitere Untersuchung auf CHiME-3-Daten durchgeführt werden.

Im Rahmen der zweiten Untersuchung werden die Signale des *isolated simulated development* Datensatzes der CHiME-3-Datenbank verwendet, der in Abschnitt 2.4 bereits beschrieben wurde. Die daraus resultierenden Bewertungsmaße ΔMOS-LQO$_{\text{GDD}}$ und ΔSNR$_{\text{GDD}}$ sind in Tab. 9.3 zusammengefasst. Außerdem sind hier SNR$_{\text{OUT}}$-Werte aus (2.41) angegeben, welche der jeweilige *a priori* SNR-Schätzer produziert hat. Um die Ergebnisse auf den CHiME-3-Daten richtig interpretieren zu können, soll daran erinnert werden, dass die verwendeten CHiME-3-Daten einen mittleren globalen eingangsseitigen SNR$_{\text{IN}}$-Wert von etwa $5.8\,$dB aufweisen. Dabei verfügen die meisten Äußerungen SNR$_{\text{IN}}$-Werte im Bereich zwischen $-2\,$dB und $14\,$dB. D. h., dass keine nennenswerte Verbesserung der Sprachsignalqualität in diesem Experiment zu erwarten ist, was die kleinen und jedoch positiven ΔMOS-LQO$_{\text{GDD}}$-Werte aus Tab. 9.3 auch bestätigen.

Folglich kann festgehalten werden, dass das GDD-Verfahren die gute Qualität des DD-Verfahrens auf den CHiME-3-Daten halten kann. Allerdings, was die Störsignaldämpfung angeht, kann das GDD-Verfahren verglichen mit dem DD-Verfahren eine höhere Leistungsfähigkeit verbuchen und zwar durchschnittlich um etwa $0.86\,$dB. Dieser Gewinn fällt dabei auch in der erwarteten Größenordnung aus, wenn man Tab. 9.2 zum Vergleich heranzieht. Ähnlich wie in der Untersuchung aus Abb. 8.5 wird auch hier eine etwas größere Steigerung der Störsignaldämpfung bei den Rauschtypen *bus* und *str* erzielt als in den Störumgebungen *caf* und *ped*. Und wenn man sich die SNR$_{\text{OUT}}$-Zahlen aus Tab. 9.3 etwas genauer anschaut, stellt man fest warum. Während das DD-Verfahren in den *caf* und *ped* Umgebungen bereits eine gute Störsignaldämpfung erzielt, hat es Schwierigkeiten, Störungen der *bus* und *str* Umgebungen zu unterdrücken. Somit zeigt das GDD-Verfahren seine Stärken genau dort, wo das DD-Verfahren seine Schwächen offenbart. Berechnet man den durchschnittlichen Gewinn des GDD-Verfahrens gegenüber dem DD-Verfahren hinsichtlich der Störsignaldämpfung, die auf den globalen eingangsseitigen SNR$_{\text{IN}}$-Wert bezogen wird, konstatiert man eine beachtliche Verbesserung der Störsignaldämpfung von $3.84\,$dB auf $4.7\,$dB.

9.5. Zusammenfassung

In diesem Kapitel wurde das weit verbreitete *Decision-Directed* Verfahren zur *a priori* SNR-Schätzung für eine Anwendung im Bereich der generalisierten SNR-Größen entwickelt, welche durch eine Einführung eines verallgemeinerten Exponenten ρ bei den spektralen Leistungen definiert sind. Dadurch entstand das generalisierte DD-Verfahren, das außer einer klassischen Gewichtungsgleichung zwei zusätzliche Berechnungsschritte für eine Hin- und Rücktransformation jeweils in den Bereich der generalisierten SNR-Größen und aus diesem Bereich heraus beinhaltet. Für die Realisierung der Rücktransformation stehen dabei zwei Möglichkeiten zur Wahl in Abhängigkeit davon, ob das Ergebnis der Gewichtung im Bereich der generalisierten SNR-Größen als Parameter oder als Zufallsgröße betrachtet wird. Neben dem Potenzfaktor ρ beinhaltet das GDD-Verfahren einen weiteren Parameter – den Gewichtsfaktor α_ρ, der auch im konventionellen DD-Verfahren bereits vorhanden ist. Eine datengetriebene Parametrisierung des GDD-Verfahrens führte dazu, dass die beiden Parameter α_ρ und ρ adaptiv in Abhängigkeit vom globalen eingangsseitigen $\mathrm{SNR}_{\mathrm{IN}}$-Wert gesetzt wurden, das aus den geschätzten *a priori* SNR-Werten einfach berechnet wurde. Geeignete Adaptionsfunktionen wurden dabei so vorgeschlagen, dass das GDD-Verfahren zu entstörten Sprachsignalen mit bester Sprachsignalqualität führt, wofür von beiden Realisierungsmöglichkeiten der Rücktransformation Gebrauch gemacht wurde. Obwohl das GDD-Verfahren etwas mehr Rechenlast erzeugt als das DD-Verfahren, ist es entsprechend Abb. 9.1 immer noch als recheneffizient einzustufen.

Eine wichtige Erkenntnis der datengetriebenen Parametrisierung war die Feststellung, dass das konventionelle DD-Verfahren für die stark gestörten Sprachsignale hinsichtlich der Qualität der prozessierten Signale bereits im geeigneten Bereich der ρ-Werte ($\rho = 1$) arbeitet und von daher optimal ist. Diese Tatsache konnte allerdings auch im vorgeschlagenen GDD-Verfahren ausgenutzt werden, das für $\rho = 1$ zum konventionellen DD-Verfahren wird. Weitere experimentelle Untersuchungen zeigten, dass das eingeführte GDD-Verfahren besonders hinsichtlich der Störsignaldämpfung das konventionelle DD-Verfahren bei allen betrachteten Rauschtypen und $\mathrm{SNR}_{\mathrm{IN}}$-Werten eindeutig übertraf. Dabei ist wichtig zu erwähnen, dass dieser Gewinn nicht auf Kosten der Qualität der entstörten Sprachsignale zustande kam. Auf den CHiME-3-Daten wurde vom GDD-Verfahren im Vergleich zum DD-Verfahren eine Störsignaldämpfung um etwa $0.9\,\mathrm{dB}$ (von $3.8\,\mathrm{dB}$ auf $4.7\,\mathrm{dB}$) erhöht, ohne dabei Verluste an Sprachsignalqualität zu verzeichnen.

10. SPP-Schätzung im generalisierten SNR-Bereich

Dieses Kapitel zeigt, dass Verwendung der generalisierten *a posteriori* SNR-Größen auch für eine SPP-Schätzung vorteilhaft sein kann, die in einem System zur spektralen Sprachsignalentstörung in Abb. 2.2 als Baustein 4 eingesetzt wird. Bereits in Abschnitt 3.4 wurde erwähnt, dass viele moderne modellbasierte Verfahren zur SPP-Schätzung von den Korrelationen zwischen den benachbarten STFT-Koeffizienten eines Sprachsignals Gebrauch machen, um einen robusten und möglichst fehlerfreien SPP-Schätzer zu erhalten [Coh01, GBM08, FW10, TVHU13, FG14, MA$^+$16]. Diese Verfahren unterscheiden sich jedoch stark in der Art und Weise, wie sie diese Korrelationen in die SPP-Schätzung einfließen lassen. So verwenden Schätzer aus [GBM08, FG14, MA$^+$16] für die Berechnung von SPP im aktuellen (k, ℓ)-ten Zeit-Frequenz-Punkt nur Beobachtungen der Zeit-Frequenz-Punkte einer bestimmten begrenzten Umgebung mit signifikanten messbaren Korrelationen. Allerdings fließen die ausgewählten Beobachtungen bei allen diesen Verfahren gleichberechtigt in die Berechnung des (k, ℓ)-ten SPP-Schätzwertes ein, ohne Berücksichtigung der aus [Coh05] oder [BCH11a] bekannten Tatsache, dass die Beobachtungen der Zeit-Frequenz-Punkte, die näher zum aktuellen Zeit-Frequenz-Punkt liegen, stärker mit seiner Beobachtung korrelieren, als Beobachtungen der weiter entfernten Zeit-Frequenz-Punkte.

Die Stärke der Korrelationen beteiligter Beobachtungen kann bei einer SPP-Schätzung dadurch berücksichtigt werden, dass verschiedene Beobachtungen unterschiedlich gewichtet werden. Dieser Ansatz führt zu einem neuartigen SPP-Schätzer, der in diesem Kapitel vorgestellt werden soll. Als Beobachtungen sollen dabei die generalisierten *a posteriori* SNR-Schätzwerte verwendet werden, die gegenüber den konventionellen *a posteriori* SNR-Größen einen zusätzlichen gewinnversprechenden Freiheitsgrad mit sich mitbringen. Außerdem ermöglichen sie eine noch fehlende Untersuchung, die zeigt, ob Verwendung konventioneller *a posteriori* SNR-Größen für eine SPP-Schätzung, die im Rahmen einer einkanaligen spektralen Sprachsignalentstörung verwendet wird, optimal ist oder nicht.

Als Grundlage für die Entwicklung eines neuen SPP-Schätzers werden in Abschnitt 10.1 zunächst Zeit-Frequenz-Korrelationen der *a posteriori* SNR-Schätzwerte untersucht, die bei gestörten Sprachsignalen beobachtet werden. In Abschnitt 10.2 wird danach ein neuartiger SPP-Schätzer hergeleitet, der auf den gewichteten generalisierten *a posteriori* SNR-Größen arbeitet. Parameter dieses SPP-Schätzers werden in Abschnitt 10.3 in Abhängigkeit vom verwendeten Kompressionsfaktor bestimmt. In Abschnitt 10.4 wird ein Kompressionsfaktor in den Experimenten mit den gestörten Sprachsignalen festgelegt. Anschließend wird in Abschnitt 10.5 die Leistungsfähigkeit des vorgeschlagenen SPP-Schätzers in den experimentellen Untersuchungen auf den CHiME-3-Daten überprüft, bevor zum Schluss in Abschnitt 10.6 eine kurze Zusammenfassung gegeben wird.

10.1. Zeit-Frequenz-Korrelationen der *a posteriori* SNR-Schätzwerte

Zunächst sollen die Korrelationseigenschaften der *a posteriori* SNR-Schätzwerte $\hat{\gamma}(k,\ell)$ von Sprachsignalen längerer Zeitdauer untersucht werden. Dafür werden ungestörte Sprachsignale der TIMIT-Datenbank von weiblichen und männlichen Sprechern jeweils mit der Länge von $30\,\mathrm{s}$ zu einem einminütigen Sprachsignal $s(n)$ zusammengesetzt, das von einem stationären Störsignal $d(n)$ des Rauschtyps '*pink*' der NOISEX-92-Datenbank bei einem globalen $\mathrm{SNR}_{\mathrm{IN}}$ von $5\,\mathrm{dB}$ additiv wie in (2.1) überlagert wird. Dabei weisen die Signale eine Abtastrate von $F_s = 16\,\mathrm{kHz}$ auf. Für eine Signalverarbeitung im Zeit-Frequenz-Bereich wird das gestörte Sprachsignal $y(n)$ einer STFT aus (2.2) mit einer FFT-Länge $K = 2^{10}$ und einem Rahmenvorschub $R = 2^8$ unter Verwendung eines Hann-Analysefensters unterzogen. Aus den daraus resultierenden STFT-Koeffizienten $Y(k,\ell)$ wird anschließend ein Spektrogramm $|Y(k,\ell)|^2$ berechnet, aus dem das Rauschleistungsdichtespektrum $\lambda_D(k,\ell)$ aus (2.7) mit Hilfe des *Minimum Statistics* Verfahrens aus [Mar01] geschätzt wird. Dabei wird die Länge des Fensters für eine Minimumsuche beim MS-Verfahren zu 96 Rahmen gewählt, welche für eine effiziente Minimumsuche in acht kleinere Unterfenster mit je 12 Rahmen aufgeteilt wird. Aus dem vorliegenden Spektrogramm $|Y(k,\ell)|^2$ und dem RLDS-Schätzwert $\hat{\lambda}_D(k,\ell)$ wird anschließend mit (2.16) einen Schätzwert des *a posteriori* SNR $\hat{\gamma}(k,\ell)$ berechnet.

Da für eine SPP-Schätzung die Zeit-Frequenz-Korrelationen von Interesse sind, die während der Sprachsignalaktivität vorliegen, wird für Berechnung von Korrelationskoeffizienten zusätzlich noch eine ideale binäre Maske $\mathrm{IBM}_S(k,\ell)$ aus den vorliegenden STFT-Koeffizienten $S(k,\ell)$ und $D(k,\ell)$ wie folgt berechnet

$$\mathrm{IBM}_S(k,\ell) = \begin{cases} 1, & |S(k,\ell)| > |D(k,\ell)| \\ 0, & \text{sonst.} \end{cases} \tag{10.1}$$

Außerdem wird eine Zeit-Frequenz-Umgebung um den (k,ℓ)-ten Zeit-Frequenz-Punkt definiert, in der die Korrelationskoeffizienten berechnet werden, siehe Abb. 10.1. Dabei wird

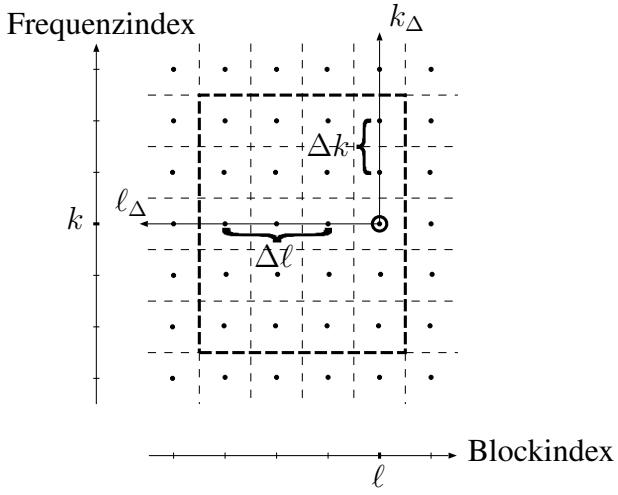


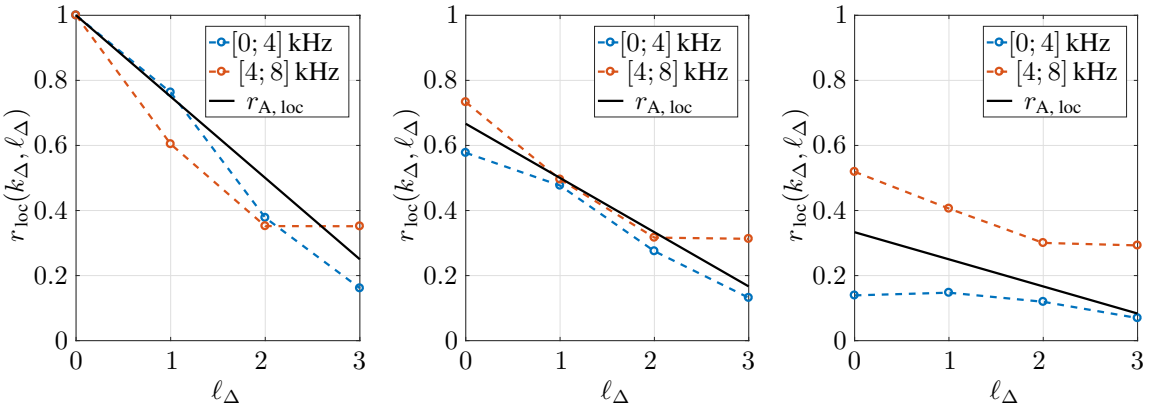
Abbildung 10.1.: Beispiel einer Zeit-Frequenz-Umgebung mit den ZF-Punkten, welche bei Berechnung von SPP im (k,ℓ)-ten ZF-Punkt verwendet werden ($\Delta k = 2$ und $\Delta\ell = 3$).

von einer ZF-Umgebung ähnlich wie in [GBM08, MHA14] gefordert, dass sie eine Kausalitätsbedingung erfüllt und nur ZF-Punkte des aktuellen Rahmens oder der vergangenen Rahmen enthält. Eine ZF-Umgebung wird durch eine Anzahl der vergangenen Rahmen $\Delta\ell$ und durch eine Anzahl der einseitigen benachbarten Frequenzbänder Δk eindeutig definiert. Somit beinhaltet eine ZF-Umgebung von jedem (k, ℓ)-ten Zeit-Frequenz-Punkt insgesamt

$$N = (2 \cdot \Delta k + 1) \cdot (\Delta\ell + 1) \tag{10.2}$$

Zeit-Frequenz-Punkte, wobei der (k, ℓ)-te Zeit-Frequenz-Punkt, der im Weiteren als ein Ankerpunkt bezeichnet wird, auch mitgezählt wird. Die Beobachtungen einer Umgebung werden bei der Berechnung der gesuchten Korrelationskoeffizienten nur dann berücksichtigt, wenn sowohl der Ankerpunkt einer Umgebung als auch mindestens $70\,\%$ aller Umgebungspunkte laut $\mathrm{IBM}_S(k, \ell)$ eine Sprachsignalaktivität aufweisen. Auf diese Weise konnten aus $\hat{\gamma}(k, \ell)$ und $\mathrm{IBM}_S(k, \ell)$ die Korrelationskoeffizienten $r(k_\Delta, \ell_\Delta)$ berechnet werden, die mit den Indizes $k_\Delta \in \{-\Delta k, \ldots, \Delta k\}$ und $\ell_\Delta \in \{0, \ldots, \Delta\ell\}$ aus Abb. 10.1 indiziert werden. Bei der Berechnung von $r(k_\Delta, \ell_\Delta)$ wird davon ausgegangen, dass die Korrelationskoeffizienten bezüglich der k_Δ-Achse symmetrisch sind $r(-k_\Delta, \ell_\Delta) = r(k_\Delta, \ell_\Delta)$. Somit müssen insgesamt nur $(\Delta k + 1) \cdot (\Delta\ell + 1)$ Korrelationskoeffizienten für $k_\Delta \geq 0$ bestimmt werden.

Korrelationseigenschaften einer lokalen Umgebung: In Abb. 10.2 sind die resultierenden Korrelationskoeffizienten $r_{\mathrm{loc}}(k_\Delta, \ell_\Delta)$ entlang der Zeitachse ℓ_Δ für verschiedene Werte von k_Δ für eine sogenannte lokale Umgebung aus Abb. 10.1 dargestellt, die mit $\Delta k_{\mathrm{loc}} = 2$ und $\Delta\ell_{\mathrm{loc}} = 3$ insgesamt $N = 20$ Zeit-Frequenz-Punkte umfasst. Dabei wird die lokale Umgebung ähnlich wie in [GBM08] definiert, sodass sie bei verwendeten Parametern F_s, K und R entlang der Zeitachse etwa $80\,\mathrm{ms}$ und entlang der Frequenzachse etwa $\pm 45\,\mathrm{Hz}$ des Spektrogramms umfasst. Um zu verdeutlichen, dass ein Sprachsignal frequenzabhängige Korrelationseigenschaften aufweist, werden mittlere Korrelationskoeffizienten in zwei verschiedenen Frequenzbereichen $[0; 4]\,\mathrm{kHz}$ und $[4; 8]\,\mathrm{kHz}$ berechnet. Betrachtet man die resultierenden Korrelationskoeffizienten $r_{\mathrm{loc}}(0, \ell_\Delta)$ entlang der Zeitachse in Abb. 10.2 (a), fällt auf, dass die nahe liegenden Zeit-Frequenz-Punkte starke Korrelationen mit dem (k, ℓ)-ten Punkt aufweisen, die allerdings mit steigenden Werten von ℓ_Δ schnell abnehmen. Dieser



(a) $k_\Delta = 0$ (b) $k_\Delta = 1$ (c) $k_\Delta = 2$

Abbildung 10.2.: Mittlere Korrelationskoeffizienten $r_{\mathrm{loc}}(k_\Delta, \ell_\Delta)$ einer lokalen Umgebung für $\Delta k_{\mathrm{loc}} = 2$ und $\Delta\ell_{\mathrm{loc}} = 3$ in verschiedenen Frequenzbereichen und eine lineare Approximation der Korrelationskoeffizienten $r_{\mathrm{A,loc}}(k_\Delta, \ell_\Delta)$: (a) $k_\Delta = 0$, (b) $k_\Delta = 1$, (c) $k_\Delta = 2$.

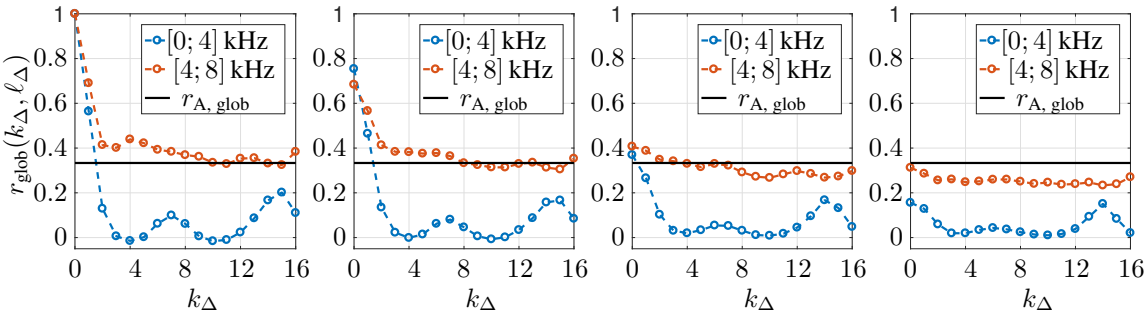
Sachverhalt ist aus den Untersuchungen in [Coh05, BCH11a] bereits bekannt. Auch in den direkt benachbarten Frequenzbändern $k+1$ und $k-1$ ist die Situation laut Abb. 10.2 (b) ähnlich, abgesehen davon, dass die Korrelationen $r_{\text{loc}}(1, \ell_\Delta)$ hier erwartungsgemäß etwas schwächer ausfallen. Entfernt man sich vom k-ten Frequenzband noch weiter, werden deutliche Unterschiede in den Korrelationseigenschaften verschiedener Frequenzbereiche bemerkbar, wie man in Abb. 10.2 (c) sieht. Während im unteren Frequenzbereich Korrelationen für alle betrachteten Werte von ℓ_Δ eigentlich vernachlässigbar sind, müssen sie im oberen Frequenzbereich immer noch berücksichtigt werden. Unterschiedliche Korrelationseigenschaften eines Sprachsignals in verschiedenen Frequenzbereichen lassen sich durch eine harmonische Struktur eines Sprachsignals erklären, die nur im unteren Frequenzbereich vorhanden ist. Die durchgeführte Analyse macht deutlich, dass die *a posteriori* SNR-Schätzwerte $\hat{\gamma}(k, \ell)$ in einer unmittelbaren lokalen Umgebung des (k, ℓ)-ten Ankerpunktes sowohl entlang Zeit als auch entlang Frequenz starke Korrelationen aufweisen, die allerdings mit der Entfernung vom Ankerpunkt in etwa linear abnehmen. Somit können die Korrelationskoeffizienten wie folgt approximiert werden:

$$r_{\text{A,loc}}(k_\Delta, \ell_\Delta) = \left(1 - \frac{|k_\Delta|}{\Delta k_{\text{loc}} + 1}\right) \cdot \left(1 - \frac{\ell_\Delta}{\Delta \ell_{\text{loc}} + 1}\right). \tag{10.3}$$

Die Korrelationskoeffizienten $r_{\text{A,lokal}}(k_\Delta, \ell_\Delta)$ werden in Abb. 10.2 neben den experimentell ermittelten Korrelationskoeffizienten dargestellt. Obwohl $r_{\text{A,loc}}(k_\Delta, \ell_\Delta)$ die Korrelationseigenschaften der *a posteriori* SNR-Schätzwerte für $k_\Delta = 2$ in unterschiedlichen Frequenzbereichen nur im Mittel gut erfassen, stellen sie sich in der restlichen lokalen Umgebung als eine durchaus brauchbare Approximation dar.

Korrelationseigenschaften einer globalen Umgebung: Während der SPP-Schätzer aus [MHA14, MA$^+$16] nur von einer lokalen Umgebung Gebrauch macht, ziehen Verfahren aus [Coh01, GBM08, FG14] für Berechnung eines SPP-Schätzwertes zusätzlich noch eine sogenannte globale Umgebung hinzu, die weit mehr Umgebungspunkte als eine lokale Umgebung beinhaltet. Die Verwendung einer globalen Umgebung sorgt im Wesentlichen dafür, dass die Unsicherheiten eines SPP-Schätzers, der auf einer lokalen Umgebung basiert, in den sicheren Sprachpausen reduziert werden [GBM08]. Während die Verwendung von einer globalen Umgebung zu Masken mit den größeren verbundenen Zeit-Frequenz-Regionen (*connected time-frequency speech presence regions* in [SA05]) führt, wird eine lokale Umgebung dafür benutzt, feinere Strukturen im Spektrogramm eines Sprachsignals zu detektieren [GBM08]. Entsprechend den Angaben in [GBM08] kann eine globale Umgebung bei den hier verwendeten Parametern F_s, K und R mit $\Delta k_{\text{glob}} = 16$ und $\Delta \ell_{\text{glob}} = 3$ definiert werden. Im Unterschied zur lokalen Umgebung erfasst die globale Umgebung einen größeren Frequenzbereich, der sich um den Ankerpunkt auf insgesamt etwa $\pm 260\,\text{Hz}$ ausdehnt.

Eine Analyse der Korrelationseigenschaften der *a posteriori* SNR-Schätzwerte in der globalen Umgebung führt zu den Korrelationskoeffizienten $r_{\text{glob}}(k_\Delta, \ell_\Delta)$, die in Abb. 10.3 entlang der Frequenzachse $k_\Delta \geq 0$ für unterschiedliche Werte von ℓ_Δ und beide Frequenzbereiche dargestellt sind. Zweifelsohne erfasst die globale Umgebung alle Korrelationen der lokalen Umgebung. Betrachtet man die Korrelationskoeffizienten der Zeit-Frequenz-Punkte, die nicht in der lokalen Umgebung vertreten sind, fällt auf, dass im unteren Frequenzbereich so gut wie keine erwähnenswerten globalen Korrelationen zu beobachten sind, was mit den Ergebnissen der Untersuchungen in [BCH11a] übereinstimmt. Dagegen sind Korrelationskoeffizienten der globalen ZF-Umgebung im oberen Frequenzbereich nicht zu unterschätzen.



(a) $\ell_\Delta = 0$ (b) $\ell_\Delta = 1$ (c) $\ell_\Delta = 2$ (d) $\ell_\Delta = 3$

Abbildung 10.3.: Mittlere Korrelationskoeffizienten $r_{\text{glob}}(k_\Delta, \ell_\Delta)$ einer globalen Umgebung für $\Delta k_{\text{glob}} = 16$ und $\Delta \ell_{\text{glob}} = 3$ in verschiedenen Frequenzbereichen und die vorgeschlagene konstante Approximation: (a) $\ell_\Delta = 0$, (b) $\ell_\Delta = 1$, (c) $\ell_\Delta = 2$, (d) $\ell_\Delta = 3$.

Allerdings können sie eher durch eine Konstante $r_{\text{A, glob}}(k_\Delta, \ell_\Delta) = r_{\text{A, glob}}$ als durch eine linear abfallende Gerade wie in (10.2) approximiert werden, wie in Abb. 10.3 verdeutlicht wird. Dabei wird $r_{\text{A, glob}}$ über eine Mittelung über alle berechneten Korrelationskoeffizienten der globalen Umgebung wie folgt berechnet:

$$r_{\text{A, glob}} = \frac{1}{N} \sum_{k_\Delta=-\Delta k}^{\Delta k} \sum_{\ell_\Delta=0}^{\Delta \ell} r_{\text{glob}}(k_\Delta, \ell_\Delta). \tag{10.4}$$

Die Analyse der Korrelationseigenschaften von $\hat{\gamma}(k, \ell)$ hier zeigt also, dass die Verwendung einer globalen Umgebung nur im oberen Frequenzbereich durch die dort vorhandenen Korrelationen gerechtfertigt ist. Im unteren Frequenzbereich dient sie eher dem Zweck, verbundene sichere Regionen im Spektrogramm zu finden, in denen mit einer relativ großen Sicherheit eine Sprachsignalaktivität vorliegt, wie in [SA05, GBM08] erläutert wird. Aus diesem Grund dürfen Beobachtungen des unteren Frequenzbereichs bei Verwendung einer globalen Umgebung ohne eine explizite Berücksichtigung ihrer Korrelationsstärke in einer SPP-Berechnung verwendet werden, was eine vereinfachte statistische Modellierung beteiligter Zufallsprozesse ermöglicht.

In einer lokalen ZF-Umgebung dagegen spielen die ZF-Korrelationen eine tragende Rolle. Um festzustellen, ob sie auch für eine lokale SPP-Schätzung gewinnbringend eingesetzt werden können, soll ein neuartiger SPP-Schätzer entwickelt werden, welcher die Korrelationsstärke einzelner Beobachtungen explizit berücksichtigen kann.

10.2. SPP-Schätzer auf gewichteten generalisierten *a posteriori* SNR-Werten

Der zu entwickelnde SPP-Schätzer, der in diesem Abschnitt hergeleitet wird, basiert auf dem SPP-Schätzer aus [GBM08], welcher die *a posteriori* SNR-Größen $\gamma(k, \ell)$ als Beobachtungen verwendet, um einen SPP-Schätzwert $\mathcal{P}(k, \ell)$ aus (2.35) zu berechnen. $\mathcal{P}(k, \ell)$ wird hier als Produkt zweier umgebungsabhängigen SPP-Schätzwerten $\mathcal{P}_{\text{loc}}(k, \ell)$ und $\mathcal{P}_{\text{glob}}(k, \ell)$ definiert, die jeweils basierend auf den Beobachtungen einer lokalen und einer globalen Umgebung berechnet werden. Dabei fließen die *a posteriori* SNR-Größen $\gamma(k, \ell)$ in die jeweilige

Berechnungsvorschrift über eine gemittelte Beobachtung gleichgewichtet ein

$$\bar{\gamma}(k,\ell) = \frac{1}{N} \sum_{k_\Delta=-\Delta k}^{\Delta k} \sum_{\ell_\Delta=0}^{\Delta \ell} \gamma(k-k_\Delta, \ell-\ell_\Delta). \tag{10.5}$$

Wie im vorigen Abschnitt diskutiert, kann eine solche Gleichgewichtung für eine globale Umgebung gerechtfertigt sein, jedoch nicht für eine lokale Umgebung. Die Korrelationsstärke verschiedener Zeit-Frequenz-Punkte einer lokalen Umgebung kann dadurch berücksichtigt werden, dass die einzelnen *a posteriori* SNR-Schätzwerte unterschiedlich gewichtet werden und für die Berechnung von $\mathcal{P}_{\text{loc}}(k,\ell)$ eine gewichtete gemittelte Beobachtung verwendet wird, die wie folgt berechnet werden kann:

$$\bar{\gamma}_w(k,\ell) = \sum_{k_\Delta=-\Delta k}^{\Delta k} \sum_{\ell_\Delta=0}^{\Delta \ell} w(k_\Delta,\ell_\Delta) \cdot \gamma(k-k_\Delta, \ell-\ell_\Delta), \tag{10.6}$$

wobei $w(k_\Delta,\ell_\Delta)$ einzelne Gewichte sind, die innerhalb einer Umgebung normiert werden

$$\sum_{k_\Delta=-\Delta k}^{\Delta k} \sum_{\ell_\Delta=0}^{\Delta \ell} w(k_\Delta,\ell_\Delta) = 1. \tag{10.7}$$

Für den zu entwickelnden SPP-Schätzer wird vorgeschlagen die Gewichte $w(k_\Delta,\ell_\Delta)$ in einen direkten Zusammenhang mit den approximierten Korrelationskoeffizienten $r_{\text{A,loc}}(k_\Delta,\ell_\Delta)$ aus (10.3) folgendermaßen zu bringen:

$$w(k_\Delta,\ell_\Delta) = \frac{r_{\text{A,loc}}(k_\Delta,\ell_\Delta)}{r_{\Sigma,\text{loc}}} \quad \text{mit} \quad r_{\Sigma,\text{loc}} = \sum_{k_\Delta=-\Delta k}^{\Delta k} \sum_{\ell_\Delta=0}^{\Delta \ell} r_{\text{A,loc}}(k_\Delta,\ell_\Delta). \tag{10.8}$$

Im Gegensatz zur Gleichgewichtung der einzelnen Beobachtungen in (10.5), wo einzelne Gewichte $w_{\text{const}}(k_\Delta,\ell_\Delta) = \frac{1}{N}$ konstant sind, nehmen die Gewichte in (10.8) mit den steigenden Werten von k_Δ und ℓ_Δ linear ab.

Um Vorteile einer Signalverarbeitung im Bereich der generalisierten *a posteriori* SNR, die zur Steigerung der Leistungsfähigkeit des *Decision-Directed* Verfahrens in Kap. 9 beitrugen, auch für eine SPP-Schätzung gewinnbringend nutzen zu können, soll beim zu entwickelnden SPP-Schätzer statt $\gamma(k,\ell)$ das generalisierte *a posteriori* SNR $\gamma_\rho(k,\ell)$ verwendet werden, das in (9.1) mit Hilfe eines Kompressionsfaktors ρ definiert ist. Dafür wird ähnlich wie in (10.6) eine mittlere gewichtete generalisierte *a posteriori* SNR-Größe definiert:

$$\bar{\gamma}_\rho(k,\ell) \triangleq \sum_{k_\Delta=-\Delta k}^{\Delta k} \sum_{\ell_\Delta=0}^{\Delta \ell} w(k_\Delta,\ell_\Delta) \cdot \gamma_\rho(k-k_\Delta, \ell-\ell_\Delta) \tag{10.9}$$

mit den linear abfallenden Gewichten aus (10.8). Man beachte, dass $\bar{\gamma}_\rho(k,\ell)$ aus (10.9) für die konstanten Gewichte $w(k_\Delta,\ell_\Delta) = w_{\text{const}}(k_\Delta,\ell_\Delta)$ und $\rho = 1$ zu $\bar{\gamma}(k,\ell)$ aus (10.5) wird.

Annahmen der statistischen Modellierung: Um den gesuchten SPP-Schätzer herzuleiten, müssen die Verteilungsdichtefunktionen $p(\bar{\gamma}_\rho(k,\ell)|H_0)$ und $p(\bar{\gamma}_\rho(k,\ell)|H_1)$ bestimmt werden, wobei H_0 und H_1 jeweils Hypothesen für eine Sprachsignalabwesenheit und für

die Sprachsignalpräsenz sind, die in (2.29) und in (2.30) bereits definiert wurden. Da bei den Herleitungen hier ähnlich wie in Abschnitt 2.1 angenommen wird, dass ein ungestörtes Sprachsignal $s(n)$ und eine Störung $d(n)$ additiv wie in (2.1) überlagert werden, gilt die Additivität der STFT-Koeffizienten (2.3) ebenfalls. Die Modellierung der STFT-Koeffizienten $S(k,\ell)$ und $D(k,\ell)$ mit (2.4) und (2.5) als zwei unkorrelierte komplexwertige mittelwertfreie normalverteilte Zufallsprozesse mit den frequenzabhängigen zeitvarianten Leistungsdichtespektren $\lambda_S(k,\ell)$ und $\lambda_D(k,\ell)$ aus (2.6) und (2.7) führt dazu, dass die STFT-Koeffizienten des gestörten Sprachsignals $Y(k,\ell)$ auch komplexwertig, mittelwertfrei und normalverteilt sind mit einer Verteilungsdichtefunktion (2.9). Bei weiteren Herleitungen wird eine Indizierung der beteiligten Zufallsgrößen der Übersichtlichkeit halber entweder tiefgesetzt oder manchmal einfach dort ausgelassen, wo dies zu keinen Missverständnissen führt.

Unter den gerade aufgeführten Annahmen unterliegen die generalisierten *a posteriori* SNR-Zufallsgrößen $\gamma_{\rho,k,\ell}|H_i$ für $i \in \{0,1\}$ eines Zeit-Frequenz-Punktes den Weibull-Verteilungen

$$p(\gamma_{\rho,k,\ell}|H_i) = \text{Weib}(\gamma_\rho;\, \lambda_{\gamma_\rho,k,\ell|H_i},\, 2\rho), \tag{10.10}$$

die in (9.4) definiert sind und sich nur in den Skalierungsfaktoren unterscheiden:

$$\lambda_{\gamma_\rho,k,\ell|H_0} = \frac{1}{\sqrt[\rho]{\Gamma(\rho+1)}}, \tag{10.11}$$

$$\lambda_{\gamma_\rho,k,\ell|H_1} = \frac{1+\xi_{k,\ell}}{\sqrt[\rho]{\Gamma(\rho+1)}}, \tag{10.12}$$

wobei $\xi(k,\ell)$ das *a priori* SNR aus (2.22) ist. Mit (9.5) resultiert dann für die Mittelwerte:

$$E[\gamma_{\rho,k,\ell}|H_0] = 1, \tag{10.13}$$

$$E[\gamma_{\rho,k,\ell}|H_1] = (1+\xi_{k,\ell})^\rho. \tag{10.14}$$

Mit dem Verschiebungssatz der Statistik und mit (9.5) ergeben sich die folgenden Varianzen:

$$\text{var}(\gamma_{\rho,k,\ell}|H_0) = c_\rho, \tag{10.15}$$

$$\text{var}(\gamma_{\rho,k,\ell}|H_1) = c_\rho \cdot (1+\xi_{k,\ell})^{2\rho} \tag{10.16}$$

mit einer vom Exponenten ρ abhängigen Konstante c_ρ, welche wie folgt ausgerechnet wird:

$$c_\rho = \frac{\Gamma(2\rho+1)}{\Gamma(\rho+1)} - 1. \tag{10.17}$$

Um $p(\bar{\gamma}_\rho(k,\ell)|H_0)$ und $p(\bar{\gamma}_\rho(k,\ell)|H_1)$ zu bestimmen, wird für die weiteren Herleitungsschritte ähnlich wie in [GBM08] angenommen, dass die generalisierten *a posteriori* SNR $\gamma_\rho(k,\ell)$ der einzelnen Zeit-Frequenz-Punkte innerhalb der lokalen Umgebung für die jeweilige Hypothese identisch verteilt sind. Für die Sprachsignalpräsenz bedeutet dies speziell, dass die beteiligten *a priori* SNR $\xi_{k,\ell}$ in der lokalen Umgebung homogen verteilt sind und in (10.12), (10.14) und (10.16) durch einen mittleren *a priori* SNR-Wert ersetzt werden

$$\bar{\xi} = \sum_{k_\Delta=-\Delta k}^{\Delta k} \sum_{\ell_\Delta=0}^{\Delta \ell} \xi_{k-k_\Delta,\ell-\ell_\Delta}. \tag{10.18}$$

Die Annahme einer identischen Verteilung scheint für hoch nichtstationäre gestörte Sprachsignale eigentlich sehr ungenau zu sein, insbesondere bei Verwendung von konstanten Gewichten. Denn selbst wenn die lokale Umgebung nur einen relativ kleinen Zeit-Frequenz-Bereich umfasst, kann es durchaus passieren, dass die vom Ankerpunkt weit entfernten Zeit-Frequenz-Punkte entweder eine Verteilung mit den stark abweichenden Skalierungsfaktoren

aufweisen oder, noch schlimmer, sogar von der alternativen Hypothese dominiert werden. An dieser Stelle wird behauptet, dass die Verwendung von den Gewichten aus (10.8), die zu den Rändern der lokalen Umgebung abfallen, die Auswirkungen dieser Annahme mildern.

Bestimmung von $p(\bar{\gamma}_{\rho,k,\ell}|H_0)$ in der Sprachsignalabwesenheit: Wird (10.13) in (10.9) für die Sprachsignalabwesenheit verwendet, resultiert für den Mittelwert mit (10.7):

$$E[\bar{\gamma}_{\rho,k,\ell}|H_0] = \sum_{k_\Delta=-\Delta k}^{\Delta k} \sum_{\ell_\Delta=0}^{\Delta \ell} w_{k_\Delta,\ell_\Delta} \cdot E[\gamma_{\rho,k-k_\Delta,\ell-\ell_\Delta}|H_0] = 1. \tag{10.19}$$

Verwendet man (10.15) in (10.9) für die Sprachsignalabwesenheit unter der Annahme der paarweisen Unkorreliertheit von $\gamma_{\rho,k-k_\Delta,\ell-\ell_\Delta}|H_0$, ergibt sich für die Varianz:

$$\text{var}(\bar{\gamma}_{\rho,k,\ell}|H_0)_{\text{unkorr}} = \sum_{k_\Delta=-\Delta k}^{\Delta k} \sum_{\ell_\Delta=0}^{\Delta \ell} w_{k_\Delta,\ell_\Delta}^2 \cdot \text{var}(\gamma_{\rho,k-k_\Delta,\ell-\ell_\Delta}|H_0) = c_\rho \cdot E_w, \tag{10.20}$$

mit einem Parameter $E_w < 1$, der aus den Quadraten der Gewichte wie folgt berechnet wird:

$$E_w = \sum_{k_\Delta=-\Delta k}^{\Delta k} \sum_{\ell_\Delta=0}^{\Delta \ell} w_{k_\Delta,\ell_\Delta}^2. \tag{10.21}$$

Allerdings sind die beteiligten Zufallsvariablen $\gamma_{\rho,k-k_\Delta,\ell-\ell_\Delta}|H_0$ aufgrund der überlappenden Signalverarbeitung einer STFT korreliert, sodass die getroffene Annahme der paarweisen Unkorreliertheit selbst bei einem weißen Störsignal $d(n)$ eigentlich falsch ist. Laut [GBM08] kann diese Unzulänglichkeit jedoch dadurch korrigiert werden, dass die in (10.20) berechnete Varianz mit Hilfe eines Korrekturfaktors c_{var} an die experimentell ermittelte Stichprobenvarianz $\hat{\text{var}}(\bar{\gamma}_{\rho,k,\ell}|H_0)$ angepasst wird, die in einem Experiment mit einem weißen Störsignal für die gegebenen Gewichte einer lokalen Umgebung aus (10.8) und für die verwendeten STFT-Parameter bestimmt werden kann. Im Unterschied zu [GBM08] hängt der Korrekturfaktor c_{var} bei der vorgeschlagenen statistischen Modellierung zusätzlich noch vom verwendeten Kompressionsfaktor ρ ab und wird experimentell folgendermaßen berechnet:

$$c_{\text{var}}(\rho) = \frac{c_\rho \cdot E_w}{\hat{\text{var}}(\bar{\gamma}_{\rho,k,\ell}|H_0)}. \tag{10.22}$$

Die konkreten Ergebnisse der experimentellen Untersuchungen bezüglich Bestimmung von $c_{\text{var}}(\rho)$ werden bei Parametrisierung des zu entwickelnden SPP-Schätzers in Abschnitt 10.3 beschrieben. Nachdem $c_{\text{var}}(\rho)$ vorliegt, kann die korrigierte Varianz von $\bar{\gamma}_\rho$ gegeben die Sprachsignalabwesenheit H_0 wie folgt berechnet werden:

$$\text{var}(\bar{\gamma}_{\rho,k,\ell}|H_0) = \frac{c_\rho \cdot E_w}{c_{\text{var}}(\rho)}. \tag{10.23}$$

Berücksichtigt man die Tatsache, dass $\bar{\gamma}_\rho$ laut (10.9) als eine Summe mehrerer gewichteter Zufallsvariablen $\gamma_{\rho,k-k_\Delta,\ell-\ell_\Delta}|H_0$ berechnet wird, darf laut dem Grenzwertsatz der Statistik unabhängig davon, welcher Verteilungsdichtefunktion die einzelnen Summanden unterliegen, angenommen werden, dass die wahre Verteilungsdichtefunktion $p(\bar{\gamma}_{\rho,k,\ell}|H_0)$ mit einer

Normalverteilung approximiert werden kann [Bri01]. Mit dem Mittelwert (10.19) und mit der Varianz (10.23) resultiert dann

$$p(\bar{\gamma}_{\rho,k,\ell}|H_0) \approx \mathcal{N}\left(\bar{\gamma}_\rho;\, 1,\, \frac{c_\rho \cdot E_w}{c_{\text{var}}(\rho)}\right). \tag{10.24}$$

Da die Normalverteilung aus (10.24) lediglich eine Approximation der wahren Verteilungsdichtefunktion $p(\bar{\gamma}_\rho|H_0)$ ist, wird auf ihre Trunkierung an der Stelle $\bar{\gamma}_\rho = 0$ verzichtet, die sonst gewährleisten könnte, dass die approximierte VDF nur für den Bereich $\bar{\gamma}_\rho > 0$ definiert ist, zumal es sich zeigte, dass die erwähnte Trunkierung keinen Einfluss auf den zu entwickelnden SPP-Schätzer hat.

Bestimmung von $p(\bar{\gamma}_{\rho,k,\ell}|H_1)$ in der Sprachsignalpräsenz: Verwendet man (10.14) in (10.9) für die Sprachsignalpräsenz und berücksichtigt die Annahme einer homogenen Verteilung der generalisierten *a posteriori* SNR der beteiligten Zeit-Frequenz-Punkte, resultiert mit (10.7) der Mittelwert:

$$\mu_{\bar{\gamma}_\rho|H_1} = E[\bar{\gamma}_{\rho,k,\ell}|H_1] = \sum_{k_\Delta=-\Delta k}^{\Delta k} \sum_{\ell_\Delta=0}^{\Delta \ell} w_{k_\Delta,\ell_\Delta} \cdot E[\gamma_{\rho,k-k_\Delta,\ell-\ell_\Delta}|H_1] = (1+\bar{\xi})^\rho. \tag{10.25}$$

Verwendet man (10.16) in (10.9) für die Sprachsignalpräsenz zunächst unter der Annahme der paarweisen Unkorreliertheit der einzelnen Zufallsvariablen $\gamma_{\rho,k-k_\Delta,\ell-\ell_\Delta}|H_1$, ergibt sich mit (10.21) die Varianz:

$$\text{var}(\bar{\gamma}_{\rho,k,\ell}|H_1)_{\text{unkorr}} = \sum_{k_\Delta=-\Delta k}^{\Delta k} \sum_{\ell_\Delta=0}^{\Delta \ell} w_{k_\Delta,\ell_\Delta}^2 \cdot \text{var}(\gamma_{\rho,k-k_\Delta,\ell-\ell_\Delta}|H_1) = c_\rho \cdot E_w \cdot (1+\bar{\xi})^{2\rho}. \tag{10.26}$$

Lässt man die Annahme der paarweisen Unkorreliertheit fallen und verwendet dafür wie in [GBM08] denselben Korrekturfaktor $c_{\text{var}}(\rho)$ aus (10.22), lautet die korrigierte Varianz für Sprachsignalpräsenz

$$\text{var}(\bar{\gamma}_{\rho,k,\ell}|H_1) = \frac{c_\rho \cdot E_w}{c_{\text{var}}(\rho)} \cdot (1+\bar{\xi})^{2\rho}. \tag{10.27}$$

Mit Verweis auf den zentralen Grenzwertsatz kann die wahre Verteilungsdichtefunktion $p(\bar{\gamma}_{\rho,k,\ell}|H_1)$ durch folgende Normalverteilung approximiert werden [Bri01]:

$$p(\bar{\gamma}_{\rho,k,\ell}|H_1) \approx \mathcal{N}\left(\bar{\gamma}_\rho;\, (1+\bar{\xi})^\rho,\, \frac{c_\rho \cdot E_w}{c_{\text{var}}(\rho)} \cdot (1+\bar{\xi})^{2\rho}\right). \tag{10.28}$$

Aus dem bereits erwähnten Grund wird bei (10.28) wieder auf eine Trunkierung der Verteilungsdichtefunktion bei $\bar{\gamma}_\rho = 0$ verzichtet.

Es ist erwähnenswert, dass $p(\bar{\gamma}_{\rho,k,\ell}|H_1)$ aus (10.28) sich für den Fall der Sprachsignalabwesenheit mit $\bar{\xi} = 0$ zu $p(\bar{\gamma}_{\rho,k,\ell}|H_0)$ aus (10.24) vereinfacht. Außerdem stellt sich die Approximation (10.28) als eine konsistente Normalverteilung (KN) dar, denn ihre Varianz

$$\text{var}(\bar{\gamma}_{\rho,k,\ell}|H_1) = \frac{c_\rho \cdot E_w}{c_{\text{var}}(\rho)} \cdot \mu_{\bar{\gamma}_\rho|H_1}^2 \tag{10.29}$$

ist kein vom Mittelwert $\mu_{\bar{\gamma}_\rho|H_1}$ unabhängiger Parameter, wie dies auch in (8.8) bei der Approximation einer Weibull-Verteilung durch eine Normalverteilung der Fall war [RSU00]. Und es gilt wieder, dass größere Varianzen mit den größeren Mittelwerten einhergehen.

Berechnung des GLR $\Lambda_{k,\ell}$ beim vorgeschlagenen SPP-Schätzer: Die berechneten Verteilungsdichtefunktionen $p(\bar{\gamma}_{\rho,k,\ell}|H_0)$ aus (10.24) und $p(\bar{\gamma}_{\rho,k,\ell}|H_1)$ aus (10.28) können jetzt in (2.36) eingesetzt werden, woraus für das generalisierte Likelihood-Verhältnis

$$\Lambda_{k,\ell}^{\text{KN}} = \frac{\Pr(H_1)}{\Pr(H_0)} \cdot \frac{1}{(1+\bar{\xi})^{\rho}} \cdot \exp\left(\frac{c_{\text{var}}}{2c_\rho E_w} \cdot \left[(\bar{\gamma}_{\rho,k,\ell} - 1)^2 - \left(\frac{\bar{\gamma}_{\rho,k,\ell}}{(1+\bar{\xi})^{\rho}} - 1 \right)^2 \right] \right). \quad (10.30)$$

resultiert. Der Übersichtlichkeit halber wird in (10.30) die Abhängigkeit des Korrekturfaktors c_{var} vom Kompressionsfaktor ρ ausgelassen.

Bevor allerdings das gerade hergeleitete GLR $\Lambda_{k,\ell}$ aus (10.30) für die Berechnung der Sprachpräsenzwahrscheinlichkeit $\mathcal{P}_{k,\ell}$ mit (2.35) verwendet werden kann, muss eine Unzulänglichkeit behoben werden, die seine Abhängigkeit von der Beobachtung $\bar{\gamma}_{\rho,k,\ell}$ betrifft und aus den vorgeschlagenen Approximationen der wahren VDF $p(\bar{\gamma}_{\rho,k,\ell}|H_0)$ und $p(\bar{\gamma}_{\rho,k,\ell}|H_1)$ durch Normalverteilungen hervorgeht. Um diesen Sachverhalt zu erklären, wird der Exponent des GLR aus (10.30) als Funktion $f(\bar{\gamma}_{\rho,k,\ell})$ wie folgt aufgefasst

$$f(\bar{\gamma}_{\rho,k,\ell}) = B \cdot (1-A) \cdot \left[(1+A) \cdot \bar{\gamma}_{\rho,k,\ell}^2 - 2\bar{\gamma}_{\rho,k,\ell} \right], \quad (10.31)$$

wobei die beiden positiven reellwertigen Konstanten A und B folgendermaßen definiert sind:

$$A = \frac{1}{(1+\bar{\xi})^{\rho}} \in (0;1) \quad \text{und} \quad B = \frac{c_{\text{var}}}{2c_\rho E_w} > 0. \quad (10.32)$$

Wie man sieht ist der Exponent $f(\bar{\gamma}_{\rho,k,\ell})$ eine quadratische Funktion der Beobachtung $\bar{\gamma}_{\rho,k,\ell}$ mit einer positiven Minimumstelle

$$\bar{\gamma}_{\rho,\text{min}} = \frac{(1+\bar{\xi})^{\rho}}{1+(1+\bar{\xi})^{\rho}}, \quad (10.33)$$

an der das generalisierte Likelihood-Verhältnis aus (10.30) seinen kleinsten Wert annimmt:

$$\Lambda_{\text{min}} = \frac{\Pr(H_1)}{\Pr(H_0)} \cdot \frac{1}{(1+\bar{\xi})^{\rho}} \cdot \exp\left(\frac{c_{\text{var}}}{2c_\rho E_w} \cdot \frac{1-(1+\bar{\xi})^{\rho}}{1+(1+\bar{\xi})^{\rho}} \right). \quad (10.34)$$

Während das GLR $\Lambda_{k,\ell}$ aus (10.30) für $\bar{\gamma}_{\rho,k,\ell} > \bar{\gamma}_{\rho,\text{min}}$ eine monoton steigende Funktion ist, stellt es sich im Bereich $\bar{\gamma}_{\rho,k,\ell} \in (0;\bar{\gamma}_{\rho,\text{min}})$ als eine monoton fallende Funktion dar. Somit ist leider auch nicht gewährleistet, dass $\mathcal{P}_{k,\ell}$ aus (2.35) eine monoton steigende Funktion der Beobachtung $\bar{\gamma}_{\rho,k,\ell}$ ist. Diese Unzulänglichkeit kann dadurch korrigiert werden, dass das GLR $\Lambda_{k,\ell}$ aus (10.30) im Bereich $\bar{\gamma}_{\rho,k,\ell} \in (0;\bar{\gamma}_{\rho,\text{min}})$ auf dem Wert Λ_{min} aus (10.34) gehalten wird, was zur folgenden Berechnungsvorschrift führt:

$$\Lambda_{k,\ell} = \begin{cases} \Lambda_{k,\ell}^{\text{KN}} \text{ aus (10.30)}, & \bar{\gamma}_\rho \geq \bar{\gamma}_{\rho,\text{min}} \\ \Lambda_{\text{min}} \text{ aus (10.34)}, & \text{sonst.} \end{cases} \quad (10.35)$$

Die in (10.35) vorgeschlagene GLR-Korrektur sorgt dafür, dass der vorgeschlagene SPP-Schätzer, der mit (2.35) berechnet wird, einen minimalen Wert $\mathcal{P}_{\text{min}}$ erhält. Dies an sich ist allerdings nicht weiter schlimm, denn auch der SPP-Schätzer aus [GBM08] weist eine Untergrenze $\mathcal{P}_{\text{min}}$ auf. Viel wichtiger ist es, dass die Verwendung von (10.35) zu einem SPP-Schätzer führt, der mit steigenden Werten der Beobachtungen $\bar{\gamma}_{\rho,k,\ell}$ nie kleinere Schätzwerte

liefert. Selbstverständlich muss dabei sichergestellt werden, dass $\mathcal{P}_{\text{min}}$ klein ist, wofür unter anderem der mittlere *a priori* SNR $\bar{\xi}$ aus (10.18) festgelegt werden muss.

Hinsichtlich der Wahl von $\bar{\xi}$ in (10.30) und (10.34) wird genauso wie in [GBM08] vorgegangen. Dabei wird $\bar{\xi}$ auf einen fixierten *a priori* SNR-Wert $\bar{\xi}_{\text{fix}}(\rho)$ gesetzt:

$$\bar{\xi}_{\text{fix}}(\rho) = \underset{\tilde{\xi}}{\text{argmin}} \int_{0.1}^{32} R(\bar{\xi}, \tilde{\xi}, \rho) d\bar{\xi}, \tag{10.36}$$

der durch eine nummerische Minimierung einer Kostenfunktion $R(\bar{\xi}, \tilde{\xi}, \rho)$ bestimmt wird. Das bestimmte Integral wird dabei wie in [GBM08] im Bereich der typischen Werte von $\bar{\xi}$ zwischen -10 dB und 15 dB ausgerechnet. Die Kostenfunktion $R(\bar{\xi}, \tilde{\xi}, \rho)$ setzt sich gleichgewichtet aus einer Falschalarmwahrscheinlichkeit $P_F(\tilde{\xi}, \rho)$ und aus einer Fehltrefferwahrscheinlichkeit $P_M(\bar{\xi}, \tilde{\xi}, \rho)$ zusammen

$$R(\bar{\xi}, \tilde{\xi}, \rho) = P_F(\tilde{\xi}, \rho) + P_M(\bar{\xi}, \tilde{\xi}, \rho), \tag{10.37}$$

wofür $\Pr(H_1) = \Pr(H_0)$ angenommen wird. Die beiden Fehlerwahrscheinlichkeiten sind wie üblich definiert:

$$P_F(\tilde{\xi}, \rho) = \int_{\bar{\gamma}_{\rho,\text{ML}}}^{\infty} p(\bar{\gamma}_\rho | H_0) d\bar{\gamma}_\rho \tag{10.38}$$

$$P_M(\bar{\xi}, \tilde{\xi}, \rho) = \int_{0}^{\bar{\gamma}_{\rho,\text{ML}}} p(\bar{\gamma}_\rho | H_1) d\bar{\gamma}_\rho. \tag{10.39}$$

Dabei ist $\tilde{\xi}$ ein *a priori* SNR-Wert, der im Laufe der durchzuführenden Minimierung durchlaufen wird und in einer ML-Entscheidungsgrenze $\bar{\gamma}_{\rho,\text{ML}}$ einfließt, die sich als ein Schnittpunkt von $p(\bar{\gamma}_\rho | H_0)$ und $p(\bar{\gamma}_\rho | H_1)$ darstellt [Kay93]. Aufgrund einer Abhängigkeit der Verteilungsdichtefunktionen $p(\bar{\gamma}_\rho | H_0)$ und $p(\bar{\gamma}_\rho | H_1)$ vom Kompressionsfaktor ρ hängt der fixierte *a priori* SNR-Wert $\bar{\xi}_{\text{fix}}(\rho)$ aus (10.36) im Unterschied zur Modellierung in [GBM08] von ρ ab. Bevor der Zusammenhang $\bar{\xi}_{\text{fix}}(\rho)$ gefunden werden kann, muss zunächst die Abhängigkeit $c_{\text{var}}(\rho)$ in den Experimenten mit einem weißen Störsignal bestimmt werden.

10.3. Parameteroptimierung des vorgeschlagenen SPP-Schätzers

Um den Zusammenhang $c_{\text{var}}(\rho)$ zu ermitteln, wird ein weißes Störsignal der gesamten Länge von 30 Sekunden aus der NOISEX-92-Datenbank verwendet. Nachdem dieses mit Hilfe einer STFT in den Zeit-Frequenz-Bereich gebracht wird, werden für jeden Zeit-Frequenz-Punkt das mittlere gewichtete generalisierte *a posteriori* SNR $\bar{\gamma}_\rho(k, \ell)$ aus (10.9) für einen bestimmten Wert von ρ mit den Gewichten $w(k_\Delta, \ell_\Delta)$ aus (10.8) berechnet, die aus den approximierten Korrelationskoeffizienten $r_{\text{A,loc}}(k_\Delta, \ell_\Delta)$ aus (10.3) einer lokalen Umgebung bestimmt werden. Da es bei dieser Untersuchung in erster Linie um Erfassung von Zeit-Frequenz-Korrelationen und nicht um Qualität einer bestimmten RLDS-Schätzung geht, wird für die Berechnung des generalisierten *a posteriori* SNR $\gamma_\rho(k, \ell)$ aus (9.1) ein RLDS-Schätzwert $\hat{\lambda}_D(k, \ell)$ verwendet, der über eine globale Mittelwertbildung über alle ZF-Punkte des vorliegenden Spektrogramms $|D(k, \ell)|^2$ bestimmt wird. Aus den so berechneten Werten $\bar{\gamma}_\rho(k, \ell)$ wird anschließend eine empirische Stichprobenvarianz $\hat{\text{var}}(\bar{\gamma}_{\rho,k,\ell} | H_0)$ bestimmt,

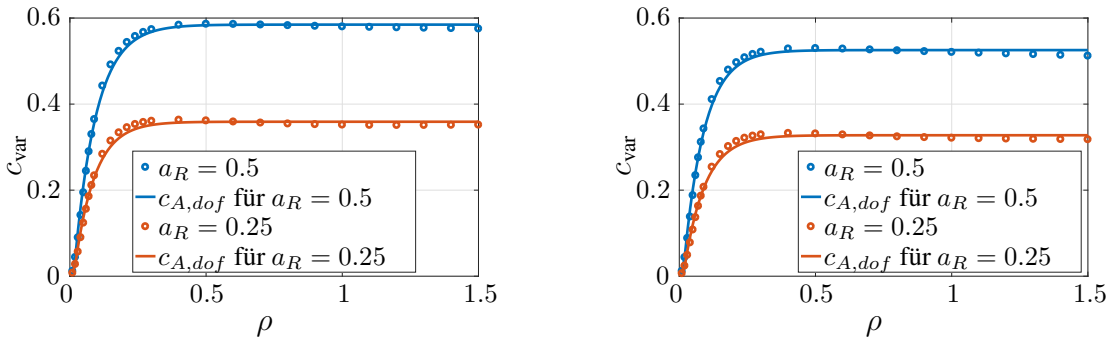
FFT-Länge, K	Verschiebungs-faktor, a_R	Δk	$\Delta \ell$	N	E_w
512	0.25	1	4	15	0.092
	0.5	1	2	9	0.146
1024	0.25	2	3	20	0.070
	0.5	2	1	10	0.130

Tabelle 10.1.: Parameter lokaler Umgebungen für verschiedene STFT-Parameter.

über welche der gesuchte Korrekturfaktor c_{var} mit (10.22) für einen bestimmten Wert von ρ ausgerechnet wird. Wiederholt man die Berechnung von c_{var} für verschiedene Werte des Kompressionsfaktors $\rho \in [0.01, 1.5]$, erhält man den gesuchten Zusammenhang $c_{\text{var}}(\rho)$.

Da verschiedene Schätzer, die in den Auswertungen dieser Arbeit verwendet wurden, für unterschiedliche STFT-Parameter entwickelt wurden, wird der vorgeschlagene SPP-Schätzer für FFT-Längen $K \in \{2^9, 2^{10}\}$ und Verschiebungsfaktoren $a_R \in \{0.25, 0.5\}$ ausgelegt. Die Parameter Δk und $\Delta \ell$ aus Abb. 10.1 werden für eine lokale Umgebung bei verschiedenen STFT-Parametern entsprechend Empfehlungen aus [GBM08] gewählt und sind in Tab. 10.1 zusammengefasst. So stimmen die Werte Δk und $\Delta \ell$ für $K = 2^9$ mit den in [GBM08] gewählten Parameterwerten überein. Außerdem sind in Tab. 10.1 die entsprechende Gesamtanzahl der Zeit-Frequenz-Punkte N aus (10.2) und resultierende Werte des Parameters E_w aus (10.21) für Gewichte aus (10.8) angegeben. Die experimentell ermittelten Zusammenhänge $c_{\text{var}}(\rho)$ für die FFT-Längen $K = \{2^9, 2^{10}\}$ und für die Verschiebungsfaktoren $a_R = \{0.25, 0.5\}$ sind in Abb. 10.4 durch die separat aufgetragenen Punkte dargestellt.

Wie man sieht, sind die Zusammenhänge $c_{\text{var}}(\rho)$ stark nichtlinear. Während Korrekturfaktoren c_{var} für kleine Werte des Kompressionsfaktors $\rho < 0.5$ mit den steigenden ρ-Werten wachsen, bleiben sie in etwa konstant für $\rho > 0.5$, als ob sie eine Sättigung erreichen. Dabei sind alle ermittelten Korrekturfaktoren deutlich kleiner als Eins ($c_{\text{var}} < 1$), wie dies auch in [GBM08] der Fall ist. Laut (10.23) bedeutet dies, dass die tatsächlichen Stichprobenvarianzen größer als die mit (10.20) berechneten Varianzen sind, die unter der Annahme der Un-



(a) $K = 2^9$ (b) $K = 2^{10}$

Abbildung 10.4.: Korrekturfaktor c_{var} als Funktion des Kompressionsfaktors ρ und verwendete Approximationen $c_{A,\text{var}}(\rho)$ für unterschiedliche FFT-Längen K und Verschiebungsfaktoren a_R: (a) $K = 2^9$, (b) $K = 2^{10}$.

FFT-Länge, K	Verschiebungs-faktor, a_R	c_1	c_2	c_3	c_{min}
512	0.25	0.36	-0.43	-13.51	$1.67 \cdot 10^{-2}$
	0.5	0.58	-0.69	-12.40	$2.78 \cdot 10^{-2}$
1024	0.25	0.33	-0.40	-13.03	$1.25 \cdot 10^{-2}$
	0.5	0.53	-0.63	-13.28	$2.50 \cdot 10^{-2}$

Tabelle 10.2.: Koeffizienten der Funktion $c_{A,\text{var}}(\rho)$ für verschiedene STFT-Parameter.

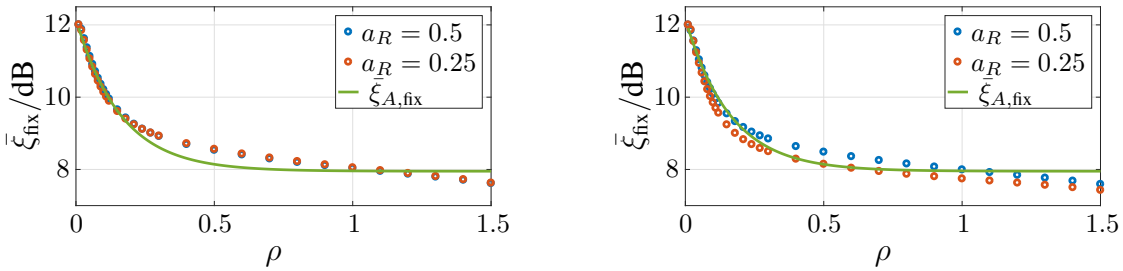
korreliertheit von generalisierten *a posteriori* SNR-Werten der benachbarten Zeit-Frequenz-Punkte berechnet werden. Angesichts der positiven Korrelationskoeffizienten einer lokalen Umgebung wie in Abb. 10.2 waren die Werte $c_{\text{var}} < 1$ auch zu erwarten. Vergleicht man die resultierenden Sättigungswerte bei verschiedenen STFT-Parametern, fällt auf, dass sie mit den sinkenden Verschiebungsfaktoren a_R deutlich kleiner werden. Das heißt, bei größeren Überlappungen der aufeinander folgenden Rahmen werden größere Korrelationen beobachtet, sodass die Varianzen aus (10.15) und (10.16) jeweils mit (10.23) und (10.27) stärker korrigiert werden müssen. Außerdem ist ein leichtes Sinken der Sättigungswerte von c_{var} zu beobachten, wenn man bei gleich bleibenden Verschiebungsfaktoren a_R die FFT-Länge K vergrößert.

Für die Parameteroptimierung des zu entwickelnden SPP-Schätzers wird vorgeschlagen, die Zusammenhänge $c_{\text{var}}(\rho)$ aus Abb. 10.4 mit Hilfe einer Funktion zu approximieren:

$$c_{A,\text{var}}(\rho) = \max(c_1 + c_2 \cdot \exp(c_3 \cdot \rho),\ c_{\text{min}}), \tag{10.40}$$

wobei c_k für $k = \{1 \ldots 3\}$ konstante Koeffizienten sind. Während die Koeffizienten c_1, c_2 und c_3 aus den experimentell ermittelten Werten von c_{var} mit Hilfe einer *Least Squares* Methode bestimmt werden, wird $c_{\text{min}} = \frac{1}{4N}$ gesetzt. Die Koeffizienten c_k und c_{min} sind für verschiedene STFT-Parameter in Tab. 10.2 zusammengefasst.

Unter Verwendung der approximierten Zusammenhänge $c_{A,\text{var}}(\rho)$ aus (10.40) mit den Koeffizienten c_k aus Tab. 10.2 können nun die Zusammenhänge $\bar{\xi}_{\text{fix}}(\rho)$ aus (10.36) ermittelt werden, wofür die entsprechende Optimierungsaufgabe nummerisch gelöst wird. Die resultierenden optimalen Werte $\bar{\xi}_{\text{fix}}$, aufgetragen in dB, sind in Abb. 10.5 als Funktion des



(a) $K = 2^9$ (b) $K = 2^{10}$

Abbildung 10.5.: Optimale fixierte *a priori* SNR $\bar{\xi}_{\text{fix}}$ als Funktion des Kompressionsfaktors ρ und die verwendete Approximationen $\bar{\xi}_{A,\text{fix}}(\rho)$ für unterschiedliche FFT-Längen K und Verschiebungsfaktoren a_R: (a) $K = 2^9$, (b) $K = 2^{10}$.

Kompressionsfaktors $\rho \in [0.01; 1.5]$ für unterschiedliche STFT-Parameter dargestellt. Beim Betrachten von Abb. 10.5 (a) lassen sich die $\bar{\xi}_{\text{fix}}$-Werte für unterschiedliche Verschiebungsfaktoren kaum voneinander unterscheiden. Für $\rho = 1$ ergibt sich ein optimaler $\bar{\xi}_{\text{fix}}$ Wert von etwa $8\,\text{dB}$, der in [GBM08] unabhängig vom Verschiebungsfaktor auch als ein optimaler Wert agiert. Wie $c_{\text{var}}(\rho)$, so ist auch der Zusammenhang $\bar{\xi}_{\text{fix}}(\rho)$ stark nichtlinear. Allerdings sinken hier die optimalen Werte von $\bar{\xi}_{\text{fix}}$ mit einem steigenden Kompressionsfaktor. Und obwohl Unterschiede in resultierenden Werten von $\bar{\xi}_{\text{fix}}$ zwischen verschiedenen Verschiebungsfaktoren in Abb. 10.5 (b) für die FFT-Länge $K = 2^{10}$ etwas größer als in Abb. 10.5 (a) ausfallen, sind sie nicht von entscheidender Bedeutung, wie weitere Simulationen zeigten. Aus diesem Grund wird für die weiteren Optimierungsschritte vorgeschlagen, den Zusammenhang $\bar{\xi}_{\text{fix}}(\rho)$ über folgende Funktion zu approximieren

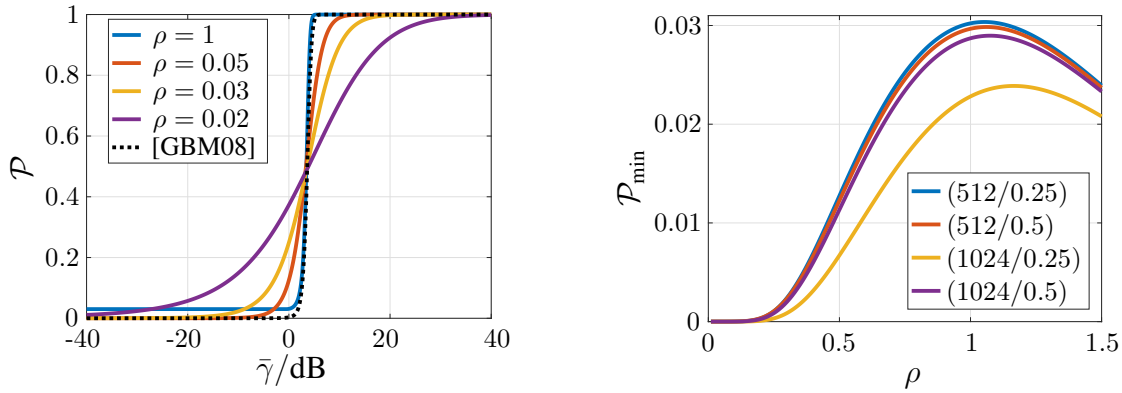
$$\bar{\xi}_{A,\text{fix}}(\rho) = \xi_1 + \xi_2 \cdot \exp(\xi_3 \cdot \rho) \tag{10.41}$$

und zwar unabhängig von den verwendeten STFT-Parametern mit einem einzigen Satz an Parametern ξ_k für $k \in \{1, 2, 3\}$. Verwendet man alle optimalen $\bar{\xi}_{\text{fix}}$ Werte, lassen sich für die Approximation $\bar{\xi}_{A,\text{fix}}(\rho)$ aus (10.41) unter Verwendung der *Least-Squares* Methode folgende Parameter bestimmen:

$$\xi_1 = 7.95,\ \xi_2 = 4.21 \text{ und } \xi_3 = \text{-}6.17. \tag{10.42}$$

Die resultierende Approximation $\bar{\xi}_{A,\text{fix}}(\rho)$ ist in Abb. 10.5 in beiden Bildern dargestellt.

Setzt man die Approximationen $c_{A,\text{var}}(\rho)$ aus (10.40) und $\bar{\xi}_{A,\text{fix}}(\rho)$ aus (10.41) in die Berechnung des generalisierten Likelihood-Verhältnisses $\Lambda_{k,\ell}$ aus (10.35) jeweils statt c_{var} und $\bar{\xi}$ ein, erhält man mit (2.35) einen SPP-Schätzer, dessen Charakteristiken über die Wahl des Kompressionsfaktors ρ gesteuert werden können. In Abb. 10.6 sind einige Eigenschaften des vorgeschlagenen SPP-Schätzers $\mathcal{P} = \Pr(H_1|\bar{\gamma})$ für $\Pr(H_1) = \Pr(H_0)$ dargestellt. In Abb. 10.6 (a) ist die resultierende *a posteriori* SPP $\mathcal{P}$, berechnet mit (10.35) als Funktion



(a) $\mathcal{P} = f(\bar{\gamma})$ (b) $\mathcal{P}_{\text{min}} = f(\rho)$

Abbildung 10.6.: Eigenschaften des vorgeschlagenen SPP-Schätzers für $\Pr(H_1) = \Pr(H_0)$: (a) $\mathcal{P} = \Pr(H_1|\bar{\gamma})$ aus (10.35) als Funktion des *a posteriori* SNR $\bar{\gamma}$ aus (10.5) unter einer idealisierten Annahme $\bar{\gamma}_\rho = \bar{\gamma}^\rho/\Gamma(\rho + 1)$ für $\rho \in \{0.05 \ldots 1.5\}$ im Vergleich zum SPP-Schätzer aus [GBM08] für $K = 2^9$ und $a_R = 0.25$. (b) Minimaler SPP-Wert $\mathcal{P}_{\text{min}}$ berechnet mit (10.34) als Funktion von ρ für verschiedene STFT-Parameter (K/a_R) in der Legende.

von $\bar{\gamma}$ aus (10.5), für einen idealisierten Fall abgebildet, dass $\bar{\gamma}_\rho = \bar{\gamma}^\rho / \Gamma(\rho + 1)$ gilt. Dieser Fall ist zwar nicht praxisrelevant, stellt sich jedoch als eine willkommene Möglichkeit die *a posteriori* SPP des vorgeschlagenen Schätzers $\Pr(H_1|\bar{\gamma})$ für verschiedene ρ-Werte gemeinsam mit $\Pr(H_1|\bar{\gamma})$ des SPP-Schätzers aus [GBM08] darzustellen. Interessanterweise liegt die Steuerungskurve für $\rho = 1$ unabhängig von einer anderen statistischen Modellierung am nächsten an den Kurven des SPP-Schätzers aus [GBM08]. Sonst fällt auf, dass durch die Wahl des Kompressionsfaktors ρ die Steilheit der Kurven gesteuert werden kann. Dabei liefert der vorgeschlagene SPP-Schätzer die relativ kleinen $\mathcal{P}_{\text{min}}$ Werte, die mit (10.34) und (2.35) berechnet werden, und nutzt somit den Wertebereich $(0; 1)$ gut aus. Dies ist eine wichtige Eigenschaft des hergeleiteten SPP-Schätzers, die allein durch die Steuerung des Parameters ξ_{fix} nicht erreicht werden kann, wie in [GBM08] gezeigt wird. Der minimale Wert $\mathcal{P}_{\text{min}}$ hängt dabei vom Kompressionsfaktor ρ ab, wie in Abb. 10.6 (b) für verschiedene STFT-Parameter dargestellt ist. Bemerkenswert ist dabei, dass $\mathcal{P}_{\text{min}}$ für kleine Werte im Bereich $\rho < 0.25$ sehr kleine Werte annimmt.

Nach der vorgestellten Parametrisierung des vorgeschlagenen SPP-Schätzers und Diskussion seiner Eigenschaften stellt sich eine wichtige Frage: welche der Steuerungskurven aus Abb. 10.6 (a) führt zur besten Leistungsfähigkeit eines Systems zur spektralen Sprachsignalentstörung? Diese entscheidende Frage soll in den Experimenten mit den gestörten Sprachsignalen beantwortet werden, die im nächsten Kapitel vorgestellt werden.

10.4. Datengetriebene Fixierung des Kompressionsfaktors

Der fixierte Kompressionsfaktor ρ_{fix} wird in den Experimenten mit den ungestörten Sprachsignalen der TIMIT-Datenbank festgelegt, die mit den Störsignalen der SPIB-Datenbank gestört werden. Dabei werden die ungestörten Sprachsignalaufnahmen unterschiedlicher weiblicher und männlicher Sprecher jeweils zu den längeren Signalen der Dauer von 30 Sekunden zusammengesetzt. Diese werden mit den Störsignalen der SPIB-Datenbank aller 15 Rauschtypen so additiv überlagert, dass gestörte Sprachsignale entstehen, welche ein eingangsseitiges globales SNR_{IN} aufweisen, das sukzessiv von 0 auf $30\,\text{dB}$ in $5\,\text{dB}$ Schritten erhöht wird. Alle Signale weisen dabei eine Abtastrate von $16\,\text{kHz}$ auf. Weitere Details zu den beiden verwendeten Datenbanken sind in Abschnitt 2.4 zu finden. Die gestörten Sprachsignale werden anschließend in einem System zur spektralen Entstörung wie in Abb. 2.2 verarbeitet, das alle vier Bausteine beinhaltet. Für eine STFT aus Abschnitt 2.1 werden ein Hann-Analysefenster der FFT-Länge $K = 2^9$ oder $K = 2^{10}$ mit einem Verschiebungsfaktor von entweder $a_R = 0.25$ oder $a_R = 0.5$ verwendet. Während das konventionelle *Minimum Statistics* Verfahren aus [Mar01] für die Schätzung der Rauschleistungsdichte $\lambda_D(k, \ell)$ eingesetzt wird, wird das *a priori* SNR $\xi(k, \ell)$ mit dem *Decision-Directed* Verfahren aus [EM84] geschätzt, in dem die LSA-Filterfunktion aus [EM85] zum Einsatz kommt. Dabei werden $\xi_{\text{min}} = -18\,\text{dB}$ und $G_{\text{min}} = -25\,\text{dB}$ wie in [Coh01] verwendet. Für die Berechnung der SPP $\mathcal{P}(k, \ell)$, welche in die Berechnung der finalen Filterfunktion wie in (2.39) als Potenz eingeht, wird neben dem vorgeschlagenen SPP-Schätzer auch das Verfahren aus [GBM08] eingesetzt. Dabei wird die Sprachpräsenzwahrscheinlichkeit $\mathcal{P}(k, \ell)$ multiplikativ ausgerechnet:

$$\mathcal{P}(k, \ell) = \mathcal{P}_{\text{lokal}}(k, \ell) \cdot \mathcal{P}_{\text{global}}(k, \ell), \tag{10.43}$$

Verschiebungs-faktor, a_R	Umgebung χ	Δk_χ	$\Delta \ell$	N_χ	$c_{\text{dof},\chi}$	$\bar{r}_\chi$	$\bar{\xi}_{\text{fix},\chi}$/dB
0.25	lokal	2	3	20	0.31	12.40	8
	global	16	3	132	0.24	63.35	3
0.5	lokal	2	1	10	0.52	10.40	8
	global	16	1	66	0.41	54.12	3

Tabelle 10.3.: Parameter des SPP-Schätzers von Gerkmann für die FFT-Länge $K = 2^{10}$.

wobei $\mathcal{P}_{\text{lokal}}(k,\ell)$ und $\mathcal{P}_{\text{global}}(k,\ell)$ jeweils unter Verwendung einer lokalen und einer globalen Umgebung berechnet werden. Basierend auf den Ergebnissen aus Abschnitt 10.1 hinsichtlich der Korrelationseigenschaften von Sprachsignalen in einer globalen Umgebung wird $\mathcal{P}_{\text{global}}(k,\ell)$ immer mit dem Verfahren aus [GBM08] berechnet, bei dem Beobachtungen gleich gewichtet werden. Die lokalen SPP-Schätzwerte $\mathcal{P}_{\text{lokal}}(k,\ell)$ werden hingegen unterschiedlich berechnet, einmal mit dem Verfahren aus [GBM08] und einmal mit dem vorgeschlagenen Verfahren, das in Abschnitt 10.2 hergeleitet und in Abschnitt 10.3 in Abhängigkeit vom Kompressionsfaktor ρ parametrisiert wurde. Der Kompressionsfaktor ρ wird dabei im Wertebereich $[0.01;\, 1]$ variiert und bleibt somit immer $\rho \leq 1$, denn weitere Untersuchungen zeigten, dass die Werte $\rho > 1$ für die Wahl von ρ_{fix} nicht relevant sind. Dafür wird zunächst der SPP-Schätzer aus [GBM08] für seinen Einsatz in einem System mit $K = 2^{10}$ parametrisiert, denn in [GBM08] wurde er nur für seine Verwendung mit $K = 2^9$ ausgelegt. Definiert man die lokale und globale Umgebungen wie in Abschnitt 10.1, können die Parameter $c_{\text{dof},\chi}$ und $\bar{\xi}_{\text{fix},\chi}$ mit $\chi \in \{\text{'lokal'}, \text{'global'}\}$ entsprechend der Vorgehensweise aus [GBM08] bestimmt werden. Die resultierenden Parameter für die beiden verwendeten Verschiebungsfaktoren $a_R \in \{0.25, 0.5\}$ werden in Tab. 10.3 zusammengefasst.

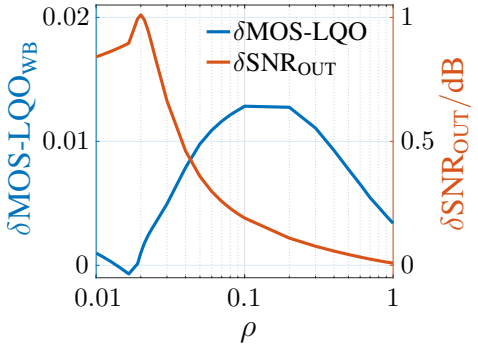
Um die beiden beschriebenen Systeme zur spektralen Sprachsignalentstörung untereinander direkt vergleichen zu können, werden zwei differenzielle Bewertungsmaße δMOS-LQO und $\delta\text{SNR}_{\text{OUT}}$ definiert:

$$\delta\text{MOS-LQO}_{\text{WB}} = \text{MOS-LQO}_{\text{WB},2} - \text{MOS-LQO}_{\text{WB},1} \tag{10.44}$$

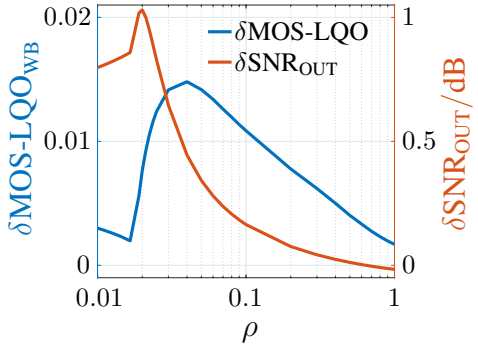
$$\delta\text{SNR}_{\text{OUT}} = \text{SNR}_{\text{OUT},2} - \text{SNR}_{\text{OUT},1}, \tag{10.45}$$

wobei $\text{MOS-LQO}_{\text{WB},i}$ und $\text{SNR}_{\text{OUT},i}$ für $i \in \{1, 2\}$ jeweils ein breitbandiges MOS-LQO-Maß und ein ausgangsseitiges globales SNR_{OUT}-Maß, die in Abschnitt 2.3 bereits eingeführt wurden. Während $i = 1$ für das System mit dem SPP-Schätzer aus [GBM08] steht, wird $i = 2$ für das System verwendet, in dem $\mathcal{P}_{\text{global}}(k,\ell)$ mit dem Verfahren aus [GBM08] und $\mathcal{P}_{\text{lokal}}(k,\ell)$ mit dem vorgeschlagenen Verfahren berechnet wurde. Die beiden Bewertungsmaße δMOS-LQO und $\delta\text{SNR}_{\text{OUT}}$ messen also die Leistungssteigerung des verwendeten Systems zur spektralen Sprachsignalentstörung, welche durch den Einsatz des vorgeschlagenen SPP-Schätzers an Stelle des Verfahrens aus [GBM08] in einer lokalen Umgebung zustande kommt.

Die resultierenden Bewertungsmaße, gemittelt über Signale aller Sprecher, über alle 15 Rauschtypen der SPIB-Datenbank und über alle simulierten SNR_{IN}-Werte, sind in Abb. 10.7 als Funktion des Kompressionsfaktors ρ für unterschiedliche STFT-Parameter dargestellt.



(a) $K = 2^9, a_R = 0.5$



(b) $K = 2^{10}, a_R = 0.25$

Abbildung 10.7.: Untersuchungen zur Wahl des fixierten Kompressionsfaktors ρ_{fix} beim entwickelten SPP-Schätzer auf den TIMIT- und SPIB-Daten für verschiedene STFT-Parameter: (a) $K = 2^9, a_R = 0.5$, (b) $K = 2^{10}, a_R = 0.25$.

Wie man sieht, zeigt der vorgeschlagene SPP-Schätzer für $\rho = 1$ in etwa dieselbe Leistungsfähigkeit wie das Verfahren aus [GBM08], unabhängig von den gewählten STFT-Parametern. Wird der Kompressionsfaktor verkleinert, verbessert sich die gesamte Leistungsfähigkeit des zweiten Systems sowohl hinsichtlich der Sprachsignalqualität als auch bezüglich der Störsignaldämpfung. Allerdings hält dieser Trend nicht bis zur unteren Intervallgrenze von ρ an, denn ab einem bestimmten Wert von ρ geht eine weitere Steigerung der Störsignaldämpfung mit dem Verlust der Sprachsignalqualität einher, wie man dies in der Regel von der spektralen Sprachsignalentstörung auch erwartet. Dabei zeigen die durchgeführten Untersuchungen, dass der vorgeschlagene SPP-Schätzer die beste Sprachsignalqualität bei den Kompressionsfaktoren im Bereich $\rho \in (0.04, 0.15)$ in Abhängigkeit der gewählten STFT-Parameter erreicht und die höchste Störsignaldämpfung für den Kompressionsfaktor von $\rho = 0.02$ unabhängig von den verwendeten STFT-Parametern.

Da die erzielten Verbesserungen der Sprachsignalqualität mit den durchschnittlichen 0.015 Punkten auf der MOS-Skala relativ klein und die erreichten Steigerungen der Störsignaldämpfung von etwa $1\,\text{dB}$ beachtenswert sind, wird für weitere Untersuchungen entschieden, den vorgeschlagenen SPP-Schätzer unabhängig von der Wahl der STFT-Parameter mit einem fixierten Kompressionsfaktor von $\rho_{\text{fix}} = 0.02$ einzusetzen. Somit müssen die verwendeten *a posteriori* SNR-Schätzwerte laut (9.9) eine relativ starke Kompression erfahren,

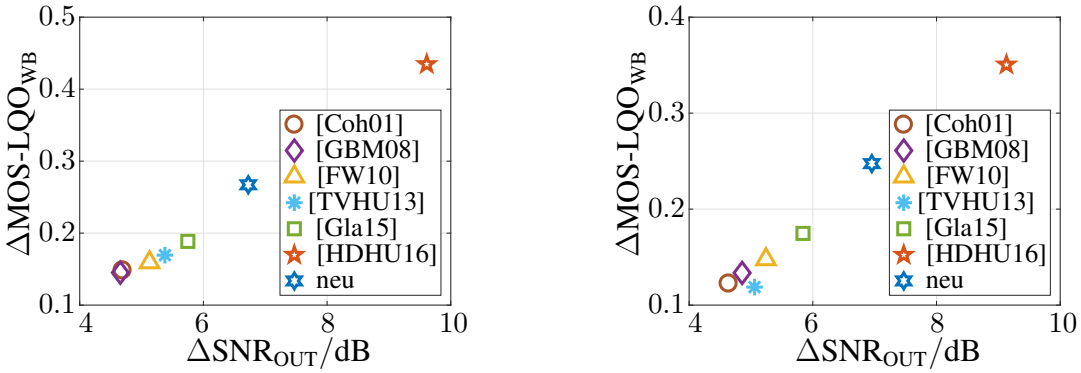
FFT-Länge, K	Verschiebungsfaktor, a_R	Δk	$\Delta \ell$	N	E_w	$c_{\text{var,fix}}$	$\bar{\xi}_{\text{fix}}/\text{dB}$
512	0.25	1	4	15	0.092	$3.2 \cdot 10^{-2}$	11.7
	0.5	1	2	9	0.146	$4.2 \cdot 10^{-2}$	11.7
1024	0.25	2	3	20	0.070	$2.2 \cdot 10^{-2}$	11.7
	0.5	2	1	10	0.130	$4.7 \cdot 10^{-2}$	11.7

Tabelle 10.4.: Fixierte Parameter des vorgeschlagenen SPP-Schätzers für $\rho_{\text{fix}} = 0.02$ in einer lokalen Umgebung für verschiedene STFT-Parameter.

damit die vorgeschlagene SPP-Schätzung zur besseren Störsignaldämpfung des Gesamtsystems beitragen kann. Die resultierenden Parameter des vorgeschlagenen SPP-Schätzers für $\rho_{\text{fix}} = 0.02$ sind in Tab. 10.4 angegeben. Man beachte, dass die Korrekturkoeffizienten $c_{\text{var,fix}}$ hier größer als die minimalen Werte c_{min} aus Tab. 10.2 sind. Sie werden tatsächlich mit Hilfe der nichtlinearen Funktion (10.40) ausgerechnet.

10.5. Untersuchungen auf den CHiME-3-Daten

Nach der Fixierung des Kompressionsfaktors kann die Leistungsfähigkeit des vorgeschlagenen SPP-Schätzers mit dem Leistungsvermögen einiger anderen Verfahren zur SPP-Schätzung verglichen werden, die in Abschnitt 3.4 eingeführt wurden. Für diese finale Auswertung werden die Daten der CHiME-3-Datenbank verwendet, die für eine einkanalige Entstörungsaufgabe sehr herausfordernd sind und in Abschnitt 2.4 bereits beschrieben wurden. Dabei wird dasselbe System zur spektralen Sprachsignalentstörung eingesetzt, das in den Experimenten im vorigen Abschnitt 10.4 eingeführt wurde. Für eine SPP-Schätzung werden in diesem System neben dem in diesem Kapitel entwickelten SPP-Schätzer sechs weitere moderne Verfahren verwendet. Während fünf dieser Verfahren [Coh01, GBM08, FW10, TVHU13, Gla15] der Klasse der modellbasierten Verfahren zugeordnet werden können, macht der SPP-Schätzer aus [HDHU16] von tiefen neuronalen Netzen Gebrauch, welche zuletzt eine außerordentliche Leistungsfähigkeit hinsichtlich der SPP-Schätzung zeigten. Da dem Verfasser dieser Arbeit Realisierungen der SPP-Schätzer aus [TVHU13], [Gla15] und [HDHU16] für die FFT-Länge $K = 2^{10}$ vorlagen, wird diese bei der bevorstehenden Auswertung verwendet. Während die statistischen Modelle der Verfahren aus [TVHU13] und [Gla15] auf den TIMIT-Daten trainiert wurden, wurde der Trainingsset der CHiME-3-Datenbank zum Training der Parameter des neuronalen Netzes aus [HDHU16] verwendet. Um verschiedene Verfahren untereinander direkt vergleichen zu können, werden in diesem Abschnitt zwei differenzielle Bewertungsmaße, $\Delta\text{SNR}_{\text{OUT}}$ und $\Delta\text{MOS-LQO}_{\text{WB}}$, verwendet, die ähnlich wie in (2.42) und (8.24) definiert werden und eine Verbesserung des jeweiligen Maßes am Systemausgang bezüglich des Systemeingangs erfassen. Die resultierenden Bewertungsmaße



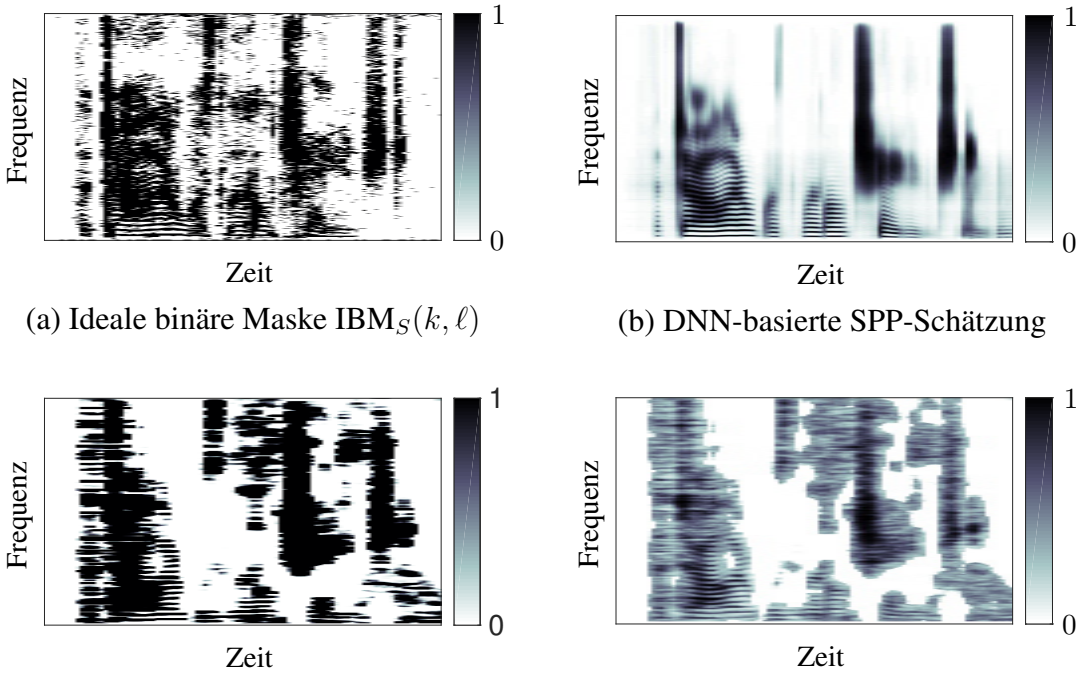
(a) $a_R = 0.25$ (b) $a_R = 0.5$

Abbildung 10.8.: Leistungsfähigkeit des entwickelten SPP-Schätzers im direkten Vergleich zu sechs anderen modernen Verfahren zur SPP-Schätzung auf CHiME-3-Daten für die FFT-Länge $K = 2^{10}$ und unterschiedliche Verschiebungsfaktoren: (a) $a_R = 0.25$, (b) $a_R = 0.5$.

dieser Untersuchung gemittelt über alle Äußerungen der CHiME-3-Daten sind in Abb. 10.8 für unterschiedliche Verschiebungsfaktoren dargestellt.

Wie erwartet, erreicht der robuste DNN-basierte SPP-Schätzer die beste Leistungsfähigkeit unter allen betrachteten Verfahren. Der neue SPP-Schätzer kommt zwar an das Leistungsvermögen des DNN-basierten SPP-Schätzers nicht heran, hebt sich jedoch deutlich von den konventionellen modellbasierten Verfahren ab. Und obwohl die Letzteren auf den CHiME-3-Daten in etwa ähnlich abschneiden, lässt sich der SPP-Schätzer aus [Gla15] hier leicht hervorheben. Verglichen mit den modellbasierten Verfahren führt der vorgeschlagene SPP-Schätzer sowohl zu einer leichten Verbesserung in der Sprachsignalqualität von etwa 0.1 Punkte auf der MOS-Skala (von $0.14 - 0.18$ auf etwa 0.26 Punkte) als auch zur Steigerung der Störsignaldämpfung um etwa 1-2 dB (von 5-6 dB auf etwa 7 dB). Insgesamt fällt auf, dass alle Verfahren bei Verwendung des Verschiebungsfaktors $a_R = 0.25$ etwas bessere Entstörungsergebnisse liefern als bei $a_R = 0.5$.

Die bemerkenswerte Leistungsfähigkeit des vorgeschlagenen SPP-Schätzers deutet darauf hin, dass die verwendete Gewichtung der Beobachtungen einer lokalen Umgebung mit (10.6), wo die einzelnen Gewichte entsprechend (10.8) aus den gemessenen Korrelationen der beteiligten Zeit-Frequenz-Punkte berechnet werden, sehr vorteilhaft für die lokale SPP-Schätzung zu sein scheint. Um dies zu verdeutlichen, werden in Abb. 10.9 die berechneten SPP-Schätzwerte $\mathcal{P}(k, \ell)$ für die Äußerung der CHiME-3-Datenbank mit dem Kennziffer *'f04_053c010p_str'* neben der dazugehörigen $\text{IBM}_S(k, \ell)$ aus (10.1) dargestellt. Während die SPP-Schätzung des besten DNN-basierten Verfahrens aus [HDHU16] in Abb. 10.9 (b) zu sehen ist, werden in Abb. 10.9 (c) und Abb. 10.9 (d) jeweils die SPP-Schätzwerte des Verfahrens aus [GBM08] und des vorgeschlagenen Verfahrens abgebildet. Aus den beiden



(a) Ideale binäre Maske $\text{IBM}_S(k, \ell)$

(b) DNN-basierte SPP-Schätzung

(c) SPP-Schätzer aus [GBM08]

(d) Vorgeschlagener SPP-Schätzer

Abbildung 10.9.: Beispiel einer SPP-Schätzung verschiedener Verfahren auf einer CHiME-3-Äußerung *'f04_053c010p_str'* für $K = 2^{10}$ und $a_R = 0.25$.

letzten Bildern lässt sich gut erahnen, dass die beiden SPP-Schätzer dieselbe globale SPP-Schätzung $\mathcal{P}_{\text{global}}(k,\ell)$ verwenden. Allerdings unterschieden sie sich sehr stark in der lokalen SPP-Schätzung $\mathcal{P}_{\text{lokal}}(k,\ell)$. Ein starker schwarz-weißer Kontrast in Abb. 10.9 (c), der auch in [GBM08] beobachtet werden kann, resultiert aus einem schmalen Transitionsbereich der Steuerungskurve dieses SPP-Schätzers, siehe Abb. 10.6 (a). Im Unterschied dazu weist die Steuerungskurve des vorgeschlagenen SPP-Schätzers mit einem kleinen fixierten Kompressionsfaktor $\rho_{\text{fix}} = 0.02$ einen viel breiteren Transitionsbereich auf und ermöglicht somit eine vielfältige Palette an resultierenden Werten von $\mathcal{P}_{\text{lokal}}(k,\ell)$. Dies ist besonders in den Zeit-Frequenz-Bereichen mit den Vokalen vorteilhaft, welche im unteren Frequenzbereich typische harmonische Tonkomplexe aufweisen. Wie man in Abb. 10.9 (d) sieht, gelingt es dem vorgeschlagenen Schätzer im Unterschied zum SPP-Schätzer aus [GBM08], solche Tonkomplexe im verrauschten Spektrogramm zu finden. Somit liefert das neue Verfahren SPP-Schätzwerte, welche ähnliche harmonische Strukturen enthalten, die beim DNN-basierten Schätzer aus Abb. 10.9 (b) vorhanden sind.

10.6. Zusammenfassung

Inspiriert vom Erfolg der spektralen Sprachsignalentstörung im Bereich der generalisierten *a posteriori* SNR aus Kap. 8 und Kap. 9 wurde in diesem Teil ein neuartiger SPP-Schätzer entwickelt, welcher auf den gewichteten generalisierten *a posteriori* SNR-Schätzwerten arbeitet. Der hier vorgeschlagene SPP-Schätzer kann dabei als Weiterentwicklung des Verfahrens aus [GBM08] angesehen werden, welche speziell die SPP-Schätzung in der lokalen Umgebung verbessert, und zwar mit Hilfe von zwei wesentlichen Aspekten - einer Gewichtung und einer Generalisierung des konventionellen *a posteriori* SNR. So werden die Beobachtungen einer lokalen Umgebung anders als im Verfahren aus [GBM08] unterschiedlich gewichtet, je nach Stärke ihrer Korrelationen mit dem aktuellen Zeit-Frequenz-Punkt, in dem die SPP geschätzt wird. Laut den durchgeführten Untersuchungen nehmen Korrelationskoeffizienten der Beobachtungen in der lokalen Umgebung mit dem Abstand zum aktuellen Zeit-Frequenz-Punkt linear ab. Aus diesem Grund wurde vorgeschlagen, die generalisierten *a posteriori* SNR-Schätzwerte vor dem Aufsummieren mit den linear abfallenden Gewichten zu multiplizieren. Die Generalisierung des *a posteriori* SNR brachte mit einem frei wählbaren Kompressionsfaktor ρ einen zusätzlichen Freiheitsgrad mit, welcher eine Steuerung der Breite des Transitionsbereichs beim vorgeschlagenen SPP-Schätzer ermöglicht.

In den Experimenten mit den Sprachsignalen der TIMIT-Datenbank, die mit verschiedenen Störsignalen der SPIB-Datenbank bei unterschiedlichen Werten des globalen eingangsseitigen SNR_{IN} überlagert wurden, wurde der neue SPP-Schätzer so parametrisiert, dass er, eingesetzt in einem System zur spektralen Sprachsignalentstörung, zur besten Leistungsfähigkeit beitrug. Dies geschah durch die Fixierung des Kompressionsfaktors, der entsprechend den durchgeführten Experimenten zum optimalen Wert von $\rho_{\text{fix}} = 0.02$ gewählt wurde. Die finale experimentelle Untersuchung auf den CHiME-3-Daten zeigte, dass der neue SPP-Schätzer alle betrachteten modellbasierten Verfahren deutlich übertrifft, da er im Stande ist, in den stark gestörten Sprachsignalen die harmonischen Tonkomponenten der Vokale zu finden. Verglichen mit den modellbasierten Verfahren führte der vorgeschlagene SPP-Schätzer sowohl zu einer leichten Verbesserung in der Sprachsignalqualität von etwa 0.1 Punkte auf der MOS-Skala (von $0.14 - 0.18$ auf etwa 0.26 Punkte) als auch zur Steigerung

der Störsignaldämpfung um etwa 1-2 dB (von 5-6 dB auf etwa 7 dB).

Sonst ist zu erwähnen, dass der in den Abschnitten 10.2 und 10.3 vorgestellte SPP-Schätzer in einer etwas abgewandelten Form auch im Rahmen einer Bachelorarbeit [Wol17] entwickelt wurde, die vom Autor dieser Arbeit ins Leben gerufen und auch betreut wurde. Im Unterschied zur vorliegenden Arbeit wurden allerdings in [Wol17] die linear abfallenden Gewichte wie in (10.8) sowohl für eine lokale als auch eine globale SPP-Schätzung verwendet, da Zeit-Frequenz-Korrelationen der *a posteriori* SNR-Schätzwerte noch nicht genauer untersucht wurden. Diese Unzulänglichkeit führte zur reduzierten Leistungsfähigkeit des entwickelten SPP-Schätzers, der das Verfahren aus [GBM08] nur leicht übertraf. Auch Verwendung der gleich gewichteten generalisierten *a posteriori* SNR-Schätzwerten wie in (10.5) angewandt gleichzeitig auf beiden Arten der Zeit-Frequenz-Umgebung, resultierte im begrenzten Leistungsvermögen des hergeleiteten SPP-Schätzers.

11. Zusammenfassung

Digitale Sprachsignalverarbeitung ist ein anerkanntes Gebiet der modernen Informationstechnik, angesiedelt in den Ingenieurwissenschaften, mit einer langjährigen Geschichte. Zu einem ihren etablierten Teilgebieten gehört die spektrale Entstörung von digitalen einkanaligen Sprachsignalen, die aufgrund einiger ungelöster Fragestellungen immer noch Gegenstand der aktuellen Forschung ist. So wird unter anderem nach den ausgeklügelten Verfahren gesucht, welche das Restrauschen im prozessierten Sprachsignal möglichst gut unterdrücken, ohne dabei solche störenden Artefakte wie *musical tones* zu produzieren, die oft mit einer fehlerhaften Schätzung von Statistiken eines Störsignals einhergehen. Eine kontrollierte Zulassung eines gewissen Rauschpegels, welches die akustische Wahrnehmung solcher Artefakte mildert, möchte man hierbei nach Möglichkeit vermeiden. Neben der Verwendung von tiefen neuronalen Netzen, werden dafür auch modellbasierte Verfahren entwickelt, die auf einer statistischen Modellierung der beteiligten Zufallsprozesse aufbauen und im Zeit-Frequenz-Bereich arbeiten. Solche Algorithmen werden als Bausteine in einem Analyse-Modifikation-Synthese System eingesetzt, in dem häufig eine recheneffiziente STFT/ISTFT Transformationspaar verwendet wird, um gestörte Sprachsignale für eine Verarbeitung in den Zeit-Frequenz-Bereich zu bringen und danach wieder zurück in den Zeitbereich. In einem solchen System kommen in der Regel folgende vier Grundbausteine zum Einsatz:

- Baustein 1 - ein Schätzer spektraler Rauschleistungsdichte,
- Baustein 2 - eine spektrale Filterfunktion für Entstörung spektraler Amplituden,
- Baustein 3 - ein *a priori* SNR-Schätzer und
- Baustein 4 - ein Schätzer der Sprachpräsenzwahrscheinlichkeit.

Um einen umfassenden Einblick in die Funktionsweise solcher Bausteine zu gewinnen, wurden in Teil A dieser Arbeit die bereits existierenden Verfahren detailliert betrachtet. Ein besonderes Augenmerk wurde dabei auf Techniken gelegt, die in diesen Verfahren zum Einsatz kommen. Außerdem wurde hier auch Verwendung von künstlichen neuronalen Netzen betrachtet, die in den letzten Jahren in der spektralen Sprachsignalentstörung stark an Bedeutung gewonnen haben. Außerdem wurden Datenbanken und Bewertungsmaße der spektralen Sprachsignalentstörung beschrieben, die in den Untersuchungen dieser Arbeit verwendet wurden. In den Teilen B und C dieser Arbeit wurden sechs verschiedene Verfahren vorgestellt, die als Bausteine in einem System zur spektralen Sprachsignalentstörung eingesetzt werden können. In Teil B wurden zunächst zwei RLDS-Schätzer entwickelt, die jeweils als Baustein 1 eingesetzt werden können. Anschließend wurde ein Bayesscher Postprozessor präsentiert, dessen Aufgabe darin besteht, eine RLDS-Schätzung seiner Vorstufe zu präzisieren. In Teil C wurden drei Verfahren zur generalisierten spektralen Sprachsignalentstörung entwickelt, die jeweils als Bausteine 2, 3 und 4 agieren können. Im Weiteren werden abschließende Zusammenfassungen aller sechs entwickelten Verfahren gegeben.

1. Alternative Steuerungsfunktion für das *Minimum Statistics* Verfahren: Im ersten Beitrag dieser Arbeit in Kap. 5 wurde eine alternative Steuerungsfunktion für das *Minimum Statistics* Verfahren aus [Mar01] vorgeschlagen, das in den Systemen zur spektralen Sprachsignalentstörung für RLDS-Schätzung sehr häufig eingesetzt wird. Speziell wurde hier die verwendete Steuerungsfunktion des zeitvarianten Glättungsfaktors kritisch untersucht, welcher im MS-Verfahren in einer rekursiven Glättung momentaner spektraler Leistungen eines gestörten Sprachsignals verwendet wird. Obwohl diese Steuerungsfunktion im MSE Sinne optimal gewählt wird, stellte sich immer noch die Frage, ob sie auch hinsichtlich solche Bewertungsmaße wie Sprachsignalqualität und Verständlichkeit die bestmögliche Wahl ist oder ob es diesbezüglich bessere Alternativen gibt. Nachdem die Vor- und Nachteile der konventionellen Steuerungsfunktion diskutiert wurden, wurde eine alternative Steuerungsfunktion vorgeschlagen. Diese beruht auf einer Bayesschen Schätzung der spektralen Rauschleistungsdichte, die im Unterschied zum MS-Verfahren nicht als Parameter sondern als Zufallsvariable modelliert wird, die einer skalierten inversen Chi-Quadrat-Verteilung mit einem Freiheitsgrad-Parameter ν unterliegt, der mit Hilfe einer empirisch gewählten Funktion gesteuert wird. Die neue Steuerungsfunktion enthält dabei einen frei wählbaren Parameter $\Delta\nu$, der im Rahmen einer experimentellen Optimierung auf den gestörten Sprachdaten zu $\Delta\nu$ = -0.3 so festgelegt wird, dass er im Mittel zu den entstörten Signalen mit bester Sprachsignalverständlichkeit gemessen mit dem STOI-Maß führt.

Laut den durchgeführten Untersuchungen verhinderte die Verwendung der neuen Steuerungsfunktion mangelnde Verfolgungsfähigkeit der rekursiven Glättung mit der konventionellen Steuerungsfunktion besonders in den Zeit-Frequenz-Bereichen, die vom Störsignal dominiert werden. In einer experimentellen Untersuchung auf den TIMIT-Daten, die von den verschiedenen Störsignalen der SPIB-Datenbank gestört wurden, zeigte sich, dass die neue Steuerungsfunktion eine Überschätzung der RLDS-Referenz verhinderte und dadurch eine leichte Verbesserung der Sprachsignalverständlichkeit bewirkte. Allerdings erwies sich die konventionelle Steuerungsfunktion als die bessere Wahl hinsichtlich der Sprachsignalqualität wie mit dem MOS-LQO$_{\text{WB}}$-Maß gemessen. Auch in der finalen Untersuchung auf den CHiME-3-Daten führte die alternative Steuerungsfunktion zu den prozessierten Sprachsignalen mit bester Sprachsignalverständlichkeit. Außerdem zeigte sich, dass eine Erhöhung des Parameters auf $\Delta\nu = 2$ zu einer vergleichbaren Verbesserung der Sprachsignalqualität führt, die von der konventionellen Steuerungsfunktion erzielt wird. Jedoch muss man dabei Verluste in der Verständlichkeit hinnehmen. Somit lässt sich durch die Wahl von $\Delta\nu$ leicht regulieren, welche Art der Verbesserung durch spektrale Sprachsignalentstörung erbracht werden soll. Ferner ergab die Untersuchung, dass eine Anpassung des Parameters α_{min} der konventionellen Steuerungsfunktion zum ähnlichen Zusammenhang zwischen Qualität und Verständlichkeit der Signale führte. Dadurch wurde ersichtlich, dass eine gleichzeitige Verbesserung der beiden Maße im Rahmen der verwendeten spektralen Entstörung nicht erreicht werden kann.

2. RLDS-Schätzung unter Verwendung eines neuronalen Netzes: Im zweiten Beitrag in Kap. 6 wurde ein robuster DNN-basierter Maskenschätzer aus [HDCHU15] in einem kausalen Schätzer einer spektralen Rauschleistungsdichte eingesetzt, welcher inspiriert von einigen modellbasierten RLDS-Schätzern auf einer rekursiven Gleichung beruht, die auch in vielen konventionellen RLDS-Schätzern verwendet wird. Dabei wird der vorgeschlagene RLDS-Schätzer vom neuronalen Netz beim Finden der Zeit-Frequenz-Punkte unterstützt, die vom Störsignal dominiert werden. Also liefert das DNN eine kausale Schätzung ent-

sprechender Wahrscheinlichkeiten und zwar allein aus dem Spektrogramm eines gestörten Sprachsignals. Wie aus der Praxis bekannt ist, stellt sich eine solche Aufgabe als sehr herausfordernd dar und ist ein großer Schwachpunkt vieler modellbasierter RLDS-Schätzer. Eingesetzt in einem System zur spektralen Sprachsignalentstörung führte der DNN-basierter RLDS-Schätzer zu einem Hybridsystem, in dem das neuronale Netz mit den konventionellen modellbasierten Verfahren kombiniert wird, die vermutlich zur besseren Generalisierungsfähigkeit des Gesamtsystems beitragen.

Um das Leistungsvermögen des vorgeschlagenen RLDS-Schätzers, dessen DNN auf den Trainingsdaten der CHiME-3-Datenbank trainiert wurde, mit Leistungsfähigkeit konventioneller modellbasierter Verfahren zur RLDS-Schätzung gerecht vergleichen zu können, wurden Parameter der Letzteren auf einem Teil der CHiME-3-Daten optimiert. Als Optimierungsmaß wurde dabei ein kombiniertes Bewertungsmaß verwendet, in welches sowohl die etablierten Bewertungsgrößen einer RLDS-Schätzung als auch die einer Sprachsignalentstörung einfließen. In einem experimentellen Vergleich zeigte sich, dass die Verwendung optimierter Parameter gemittelt über alle betrachteten Verfahren sowohl zu einer genaueren RLDS-Schätzung als auch zu einer qualitativ besseren Signalentstörung führen. Da allerdings einige Bewertungsmaße auf Kosten der anderen verbessert wurden, wurden die konventionellen modellbasierten Verfahren in einer finalen Auswertung sowohl mit dem von den jeweiligen Autoren empfohlenen als auch mit den optimierten Parametern eingesetzt. Dabei wurde gezeigt, dass der DNN-basierte RLDS-Schätzer alle modellbasierten Verfahren in allen betrachteten Bewertungsmaßen übertrifft und zwar unabhängig davon, ob bei den Letzteren ein empfohlener oder ein optimierter Parametersatz verwendet wird.

3. Bayesscher Postprozessor zur RLDS-Schätzung: Im dritten Beitrag in Kap. 7 wurde ein Bayessches Verfahren zur RLDS-Schätzung entwickelt, das in einem System zur spektralen Sprachsignalentstörung als Postprozessor agiert. Der vorgeschlagene MAP-basierte Postprozessor wurde in zwei Versionen vorgestellt, und zwar nichtoptimiert und optimiert. Ein Einsatz des nichtoptimierten MAPB-Postprozessors in der spektralen Sprachsignalentstörung zeigte, dass das vorgeschlagene Verfahren zur Reduktion der Schätzfehlervarianz der RLDS-Schätzung führt und somit eine Verbesserung der Qualität der entstörten Signale bewirkt. Eine numerische Qualitätsanalyse des nichtoptimierten MAPB-Postprozessors deckte jedoch einige Unzulänglichkeiten des Verfahrens auf, welche durch eine Biaskorrektur und eine Bandbreitenanpassung weitgehend beseitigt wurden. Dadurch entstand die optimierte Version des MAPB-Schätzers, dessen Leistungsfähigkeit in den umfangreichen Experimenten mit insgesamt zwölf verschiedenen RLDS-Schätzern in der ersten Systemstufe untersucht wurde.

Die Untersuchungen auf den Daten der TIMIT-Datenbank zeigten, dass der MAPB-Postprozessor die RLDS-Schätzung der ersten Systemstufe sowohl hinsichtlich des mittleren logarithmischen Schätzfehlers als auch bezüglich der mittleren logarithmischen Schätzfehlervarianz präzisieren kann, insbesondere bei solchen nichtstationären Störungen wie *babble noise*. Genauere RLDS-Schätzung sorgt dabei für eine leichte Verbesserung der Qualität der prozessierten Sprachsignale, die um so größer ausfällt, je stärker Signale verrauscht sind. Auch auf den CHiME-3-Daten wurde eine Verbesserung der Sprachsignalqualität beobachtet, die allerdings in manchen Fällen mit leichten Verlusten in der Störsignaldämpfung einherging. Die einzige Ausnahme hier stellten der DNN-basierte RLDS-Schätzer und das MMSE-BM-Verfahren dar, welche bereits eine gute Leistungsfähigkeit lieferten und vom vorgeschlagenen MAPB-Postprozessor nicht verbessert werden konnten.

4. MAP-Schätzer generalisierter log-spektraler Amplituden: Im vierten Beitrag dieser Arbeit in Kap. 8 wurde eine spektrale Filterfunktion mit Hilfe einer neuartigen Nichtlinearität hergeleitet, die im Bereich der logarithmischen generalisierten spektralen Amplituden angesiedelt ist. Dabei stellte sich die LGSA-Nichtlinearität als eine Kombination zweier bekannter Nichtlinearitäten dar, eines natürlichen Logarithmus und einer Kompressionsfunktion, welche in der generalisierten modellbasierten Sprachsignalentstörung häufig zum Einsatz kommt. Um eine einfache recheneffiziente Berechnungsvorschrift für die neue LGSA-Filterfunktion zu erhalten, wurden die Weibull-Verteilungen beteiligter Zufallsprozesse auf dem Weg in den LGSA-Bereich im Rahmen einer Fünf-Schritte-Herleitung so approximiert, dass am Ende die bedingte *a posteriori* Verteilungsdichtefunktion der LGSA-Koeffizienten berechnet werden konnte. Der Modalwert dieser Verteilung führte zum gesuchten recheneffizienten MAP-basierten LGSA-Schätzer. Um den verwendeten Approximationen und einigen Modellannahmen, die in der Realität verletzt werden, entgegenzuwirken, wurde anschließend eine Strategie zur Erhöhung der Flexibilität der statistischen Modellierung vorgeschlagen. Als Nebenprodukt entstand dabei eine modifizierte Version des generalisierten MMSE-basierten PGSS-Schätzers aus [STCT98]. Anschließend folgte eine datengetriebene Parametrisierung der beiden Schätzer mit dem Ziel, die beste Qualität der entstörten Signale zu erreichen. Daraus resultierten zwei generalisierte spektrale Filterfunktionen: der bekannte modifizierte MMSE-PGSS-Schätzer und der vorgeschlagene MAP-LGSA-Schätzer.

Ein experimenteller Vergleich der beiden Filterfunktionen mit dem MMSE-LSA-Schätzer aus [EM85] offenbarte eine hervorragende Leistungsfähigkeit des MAP-basierten LGSA-Schätzers, der im Rahmen einer einkanaligen Sprachsignalentstörung der CHiME-3-Daten den MMSE-LSA-Schätzer hinsichtlich der Störsignaldämpfung im Mittel um etwa $1.4\,\mathrm{dB}$ (von $4.2\,\mathrm{dB}$ auf $5.6\,\mathrm{dB}$) übertraf und dabei keine nennenswerten Verluste in der Qualität der prozessierten Sprachsignale verursachte. Somit wurde gezeigt, dass der vorgeschlagene MAP-LGSA-Schätzer den Kompromiss zwischen der Sprachsignalqualität und der Störsignaldämpfung etwas besser löst als der MMSE-LSA- und der MMSE-PGSS-Schätzer.

5. Generalisiertes *Decision-Directed* Verfahren: Im fünften Beitrag in Kap. 9 wurde das weit verbreitete *Decision-Directed* Verfahren zur *a priori* SNR-Schätzung für eine Anwendung im Bereich der generalisierten SNR-Größen entwickelt, welche durch eine Einführung eines verallgemeinerten Exponenten ρ bei den spektralen Leistungen definiert sind. Dadurch entstand das generalisierte DD-Verfahren, das außer einer klassischen Gewichtungsgleichung zwei zusätzliche Berechnungsschritte für eine Hin- und Rücktransformation jeweils in den Bereich der generalisierten SNR-Größen und aus diesem Bereich heraus beinhaltet. Für die Realisierung der Rücktransformation stehen dabei zwei Möglichkeiten zur Wahl in Abhängigkeit davon, ob das Ergebnis der Gewichtung im Bereich der generalisierten SNR-Größen als Parameter oder als Zufallsgröße betrachtet wird. Neben dem Potenzfaktor ρ beinhaltet das GDD-Verfahren einen weiteren Parameter – den Gewichtsfaktor α_ρ, der auch im konventionellen DD-Verfahren bereits vorhanden ist. Eine datengetriebene Parametrisierung des GDD-Verfahrens führte dazu, dass die beiden Parameter α_ρ und ρ adaptiv in Abhängigkeit vom globalen eingangsseitigen SNR-Wert gesetzt wurden, das aus den geschätzten *a priori* SNR-Werten berechnet wurde. Geeignete Adaptionsfunktionen wurden dabei so vorgeschlagen, dass das GDD-Verfahren zu entstörten Sprachsignalen mit bester Sprachsignalqualität führt, wofür von beiden Realisierungsmöglichkeiten der Rücktransformation Gebrauch gemacht wurde. Obwohl das GDD-Verfahren etwas mehr Rechenlast erzeugt als das DD-Verfahren, ist es immer noch als recheneffizient einzustufen.

Eine wichtige Erkenntnis der datengetriebenen Parametrisierung war die Feststellung, dass das konventionelle DD-Verfahren für die stark gestörten Sprachsignale hinsichtlich der Qualität der prozessierten Signale bereits im geeigneten Bereich der ρ-Werte ($\rho = 1$) arbeitet und von daher optimal ist. Diese Tatsache konnte allerdings auch im vorgeschlagenen GDD-Verfahren ausgenutzt werden, das für $\rho = 1$ zum konventionellen DD-Verfahren wird. Weitere experimentelle Untersuchungen zeigten, dass das eingeführte GDD-Verfahren besonders hinsichtlich der Störsignaldämpfung das konventionelle DD-Verfahren bei allen betrachteten Rauschtypen und $\mathrm{SNR}_{\mathrm{IN}}$-Werten eindeutig übertraf. Dabei ist wichtig zu erwähnen, dass dieser Gewinn nicht auf Kosten der Qualität der entstörten Sprachsignale zustande kam. Auf den CHiME-3-Daten wurde vom GDD-Verfahren im Vergleich zum DD-Verfahren eine Störsignaldämpfung um etwa $0.9\,\mathrm{dB}$ (von $3.8\,\mathrm{dB}$ auf $4.7\,\mathrm{dB}$) erhöht, ohne dabei Verluste an Sprachsignalqualität zu verzeichnen.

6. SPP-Schätzung im generalisierten SNR-Bereich: Im sechsten Beitrag in Kap. 10 wurde ein neuartiger SPP-Schätzer entwickelt, welcher auf den gewichteten generalisierten *a posteriori* SNR-Schätzwerten arbeitet. Der hier vorgeschlagene SPP-Schätzer kann dabei als Weiterentwicklung des Verfahrens aus [GBM08] angesehen werden, welche speziell die SPP-Schätzung in der lokalen Umgebung verbessert, und zwar mit Hilfe von zwei wesentlichen Aspekten - einer Gewichtung und einer Generalisierung des konventionellen *a posteriori* SNR. So werden die Beobachtungen einer lokalen Umgebung anders als im Verfahren aus [GBM08] unterschiedlich gewichtet, je nach Stärke ihrer Korrelationen mit dem aktuellen Zeit-Frequenz-Punkt, in dem die SPP geschätzt wird. Laut den durchgeführten Untersuchungen nehmen Korrelationskoeffizienten der Beobachtungen in der lokalen Umgebung mit dem Abstand zum aktuellen Zeit-Frequenz-Punkt linear ab. Aus diesem Grund wurde vorgeschlagen, die generalisierten *a posteriori* SNR-Schätzwerte vor dem Aufsummieren mit den linear abfallenden Gewichten zu multiplizieren. Die Generalisierung des *a posteriori* SNR brachte mit einem frei wählbaren Kompressionsfaktor ρ einen zusätzlichen Freiheitsgrad mit, welcher eine Steuerung der Breite des Transitionsbereichs beim vorgeschlagenen SPP-Schätzer ermöglicht.

In den Experimenten mit den Sprachsignalen der TIMIT-Datenbank, die mit verschiedenen Störsignalen der SPIB-Datenbank bei unterschiedlichen Werten des globalen eingangsseitigen $\mathrm{SNR}_{\mathrm{IN}}$ überlagert wurden, wurde der neue SPP-Schätzer so parametrisiert, dass er, eingesetzt in einem System zur spektralen Sprachsignalentstörung, zur besten Leistungsfähigkeit beitrug. Dies geschah durch die Fixierung des Kompressionsfaktors, der entsprechend den durchgeführten Experimenten zum optimalen Wert von $\rho_{\mathrm{fix}} = 0.02$ gewählt wurde. Die finale experimentelle Untersuchung auf den CHiME-3-Daten zeigte, dass der neue SPP-Schätzer alle betrachteten modellbasierten Verfahren deutlich übertrifft, da er im Stande ist, in den stark gestörten Sprachsignalen die harmonischen Tonkomponenten der Vokale zu finden. Verglichen mit den modellbasierten Verfahren führte der vorgeschlagene SPP-Schätzer sowohl zu einer leichten Verbesserung in der Sprachsignalqualität von etwa 0.1 Punkte auf der MOS-Skala (von $0.14 - 0.18$ auf etwa 0.26 Punkte) als auch zur Steigerung der Störsignaldämpfung um etwa 1-$2\,\mathrm{dB}$ (von 5-$6\,\mathrm{dB}$ auf etwa $7\,\mathrm{dB}$).

Gemeinsame Verwendung entwickelter Verfahren für spektrale Entstörung: Zum Schluss werden die in den Kapiteln 6-10 entwickelten Verfahren in einem System zur spektralen Sprachsignalentstörung wie in Abb. 2.2 gemeinsam eingesetzt und auf den stark gestörten CHiME-3-Daten evaluiert, die in Abschnitt 2.4 beschrieben sind. Dabei werden einzelne Bausteine eines konventionellen Systems zur spektralen Sprachsignalentstörung nach

System	Baustein 1	Baustein 2	Baustein 3	Baustein 4	ΔSNR / dB	ΔMOS-LQO$_{\text{WB}}$
1.	OSMS	LSA	DD	SPP-FP	4.54	0.121
2.	OSMS	**LGSA**	DD	SPP-FP	5.75	0.160
3.	OSMS	**LGSA**	**GDD**	SPP-FP	6.50	0.171
4.	OSMS	**LGSA**	**GDD**	**SPP-GW**	8.23	0.270
5.	OSMS/**MAPB**	**LGSA**	**GDD**	**SPP-GW**	9.32	0.275
6.	**DNN-NPP**	**LGSA**	**GDD**	**SPP-GW**	8.60	0.220
7.	**DNN-NPP**	**LGSA**	**GDD**	DNN-SPP	11.51	0.228
8.	OSMS/**MAPB**	**LGSA**	**GDD**	DNN-SPP	11.46	0.340

Tabelle 11.1.: Architektur der Systeme mit den entwickelten Verfahren und ihre Leistungsfähigkeit auf den CHiME-3-Daten mit $\text{SNR}_{\text{IN}} = 5.8\,\text{dB}$ und $\text{MOS-LQO}_{\text{WB,IN}} = 1.27$.

und nach mit den in dieser Arbeit entwickelten Verfahren ersetzt[1]. Insgesamt beteiligen sich acht verschiedene Systeme an der Auswertung, deren Architekturen und die resultierende Leistungsfähigkeit in Tab. 11.1 gegeben sind. Demnach besteht das System 1 aus vier konventionellen Schätzverfahren: dem OSMS-Verfahren aus [Mar01], dem MMSE-LSA-Schätzer aus [EM85], dem DD-Verfahren aus [EM84] und dem SPP-FP-Schätzer aus [GBM08], die entsprechend Empfehlungen jeweiliger Autoren parametrisiert werden. Bei allen Systemen fließt die Schätzung von SPP $\mathcal{P}(k,\ell)$ wie in (2.39) als Potenz in die Berechnung der finalen Filterfunktion ein, wobei $G_{H_0} = -20\,\text{dB}$ und $\xi_{\text{min}} = -18\,\text{dB}$ aus (2.25) verwendet werden. Die resultierenden Bewertungsmaße ΔSNR aus (2.42) und ΔMOS-LQO aus (8.24) aus Tab. 11.1 sind außerdem in Abb. 11.1 graphisch visualisiert, wobei modellbasierte Systeme durch Kreise und Hybridsysteme (modellbasierte Verfahren kombiniert mit den DNN-basierten Verfahren) dargestellt sind.

Die durchgeführte Untersuchung resultiert in den folgenden interessanten Ergebnissen:

1. Das konventionelle System erreicht auf den CHiME-3-Daten ΔSNR von etwa $4.5\,\text{dB}$ und ΔMOS-LQO von 0.12 Punkte auf der MOS-Skala.

2. Verwendet man statt MMSE-LSA-Schätzer die MAP-LGSA-Filterfunktion aus Kap. 8, lässt sich neben einer leichten Steigerung der Sprachsignalqualität auch eine Erhöhung in Störsignaldämpfung von etwa $1.2\,\text{dB}$ feststellen.

3. Ersetzt man das DD-Verfahren zu *a priori* SNR-Schätzung durch das GDD-Verfahren aus Kap. 9, wird ein weiterer Gewinn der Störsignaldämpfung von $0.75\,\text{dB}$ erzielt.

[1] Da in dieser Untersuchung nur Störsignaldämpfung und Sprachsignalqualität untersucht werden, nimmt das BSMS-Verfahren aus Kap. 5, das für eine Verbesserung der Sprachsignalverständlichkeit entwickelt wurde, an dieser Auswertung nicht teil.

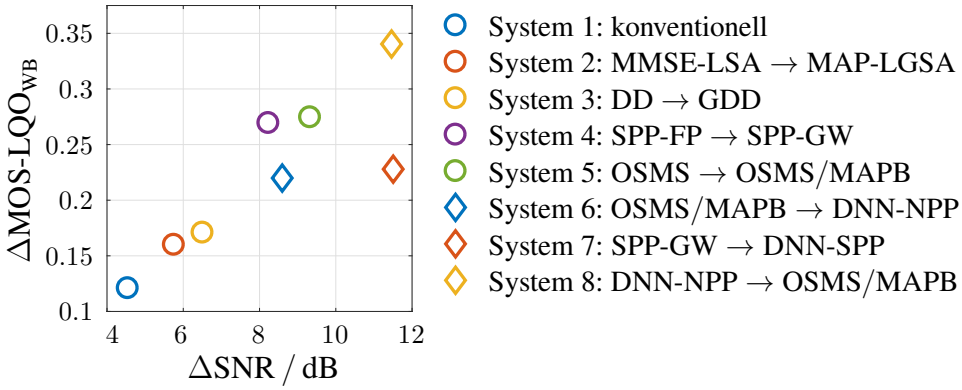


Abbildung 11.1.: Leistungsfähigkeit der Systeme zur spektralen Sprachsignalentstörung aus Abb. 2.2 auf den CHiME-3-Daten mit $\text{SNR}_{\text{IN}} = 5.8\,\text{dB}$ und $\text{MOS-LQO}_{\text{WB,IN}} = 1.27$, wenn einzelne Systembausteine nach und nach mit den entwickelten Verfahren ersetzt werden.

4. Tauscht man außerdem den SPP-FP-Schätzer durch das SPP-GW-Verfahren aus[2], das in Kap. 10 entwickelt wurde, werden sowohl Signalqualität um etwa 0.1 Punkte auf der MOS-Skala als auch Störsignaldämpfung um gute $1.7\,\text{dB}$ weiter erhöht.
5. Setzt man zusätzlich noch den optimierten MAP-basierten Postprozessor aus Kap. 7 zur Präzisierung der RLDS-Schätzung ein, ist ein weiteres leichtes Wachstum der Störsignaldämpfung von etwa $0.9\,\text{dB}$ bei gleich bleibender Signalqualität zu beobachten.

Das System 5 erzielt dabei $\Delta\text{SNR} \approx 9.3\,\text{dB}$ und $\Delta\text{MOS-LQO} \approx 0.28$. Verglichen mit dem konventionellen System 1 werden somit sowohl die Störsignaldämpfung als auch die Sprachsignalqualität der prozessierten Signale mehr als verdoppelt. Dabei fällt eine permanente Steigerung der Leistungsfähigkeit spektraler Entstörung beim Austauschen alter Systemkomponenten durch die neu entwickelten Verfahren auf. Das Ausmaß der erzielten Verbesserung wird deutlich, wenn das Leistungsvermögen der Hybridsysteme betrachtet wird.

6. Verwendet man den DNN-basierten RLDS-Schätzer aus Kap. 6 statt des OSMS-Verfahrens mit dem MAPB-Postprozessor, sinken beide Bewertungsmaße leicht.
7. Erst wenn zusätzlich noch ein DNN-basierter SPP-Schätzer als Baustein 4 verwendet wird, steigt die Störsignaldämpfung um gute $2.9\,\text{dB}$.
8. Allerdings werden sowohl die beste Sprachsignalqualität als auch die beste Störsignaldämpfung erst dann erreicht, wenn der DNN-basierte RLDS-Schätzer durch das OSMS-Verfahren mit dem MAPB-Postprozessor ausgetauscht wird.

Somit legt die durchgeführte Untersuchung nah, dass die beste Leistung nicht vom System 7 mit den meisten DNN-basierten Komponenten erreicht, sondern vom Hybridsystem 8, was vermutlich mit besserer Generalisierungsfähigkeit der Hybridsysteme einhergeht. Insofern wird ein Schwerpunkt zukünftiger Forschungsarbeiten sicherlich auf der Entwicklung solcher Hybridsysteme liegen, in denen modellbasierten Verfahren mit den tiefen neuronalen Netzen kombiniert werden, die in der modernen spektralen Sprachsignalentstörung unverzichtbar zu sein scheinen.

[2]Die Abkürzung GW steht hier für *generalized weighted*, denn beim in Kap. 10 vorgeschlagenen SPP-GW Schätzer die generalisierten *a posteriori* SNR-Größen zusätzlich gewichtet werden.

A. Anhang

Parametrisierung rekursiver Glättung erster Ordnung

Die rekursive Glättung erster Ordnung (2.48) kommt in der spektralen Signalverarbeitung sehr häufig zum Einsatz. Neben der Berechnung der RLDS Referenz wie in (2.48) wird sie auch von vielen Verfahren zur Schätzung des RLDS, des *a priori* SNR und der SPP verwendet. Die Filtereigenschaften der rekursiven Glättung, die ja ein zeit-diskreter IIR-Filter erster Ordnung ist, werden allerdings nicht allein durch die Angabe des Glättungsparameters festgelegt, dessen Referenzwert α_{ref} in den Publikationen des jeweiligen Schätzverfahrens meistens angegeben wird. Der absolute Wert von α_{ref}, der im Wertebereich $(0\,;1)$ liegt, verrät nur, dass die rekursive Glättung (2.48) ein *bounded iput bounded output* (BIBO) stabiles Tiefpassfilter ist. In die Berechnung von solchen Kennzahlen dieses Filters, wie die Grenzfrequenz (engl. *cut-off frequency*) f_c oder die Zeitkonstante τ, die jeweils in Herz und Sekunde gemessen werden, fließt neben α_{ref} noch das Verhältnis $\frac{R_{\text{ref}}}{F_{s,\text{ref}}}$ des STFT-Rahmenvorschubs R_{ref} zur Abtastrate des zeitdiskreten Signals $F_{s,\text{ref}}$ ein, die in den Experimenten der jeweiligen Publikation verwendet wurden. Dieses Verhältnis gibt den zeitlichen Abstand $\Delta T_{\text{ref}} = \frac{R_{\text{ref}}}{F_{s,\text{ref}}}$ zwischen den aufeinander folgenden Rahmen $\ell - 1$ und ℓ in Sekunden an. Möchte man ein Schätzverfahren in einem System mit einem anderen Verhältnis $\frac{R}{F_s}$ realisieren, muss der Glättungsparameter unbedingt umgerechnet werden, damit die Filtereigenschaften der rekursiven Glättung wie f_c und τ dieselben bleiben. Somit ist die Angabe von α_{ref} allein für Parametrisierung rekursiver Glättung erster Ordnung (2.48) nicht eindeutig, sodass benutzerfreundlichere Wege für eine eindeutige Parametrisierung gesucht werden müssen. Und da in der Fachliteratur solche Wege nur ansatzweise und ohne eine vollständige Einführung benutzt werden, werden sie in diesem Anhang zunächst ausführlich hergeleitet und anschließend diskutiert.

Zeitdiskretes IIR-Filter erster Ordnung: Die rekursive Glättung erster Ordnung wie in (2.48) wird für die weiteren Herleitungen in der Form einer Differenzengleichung angegeben

$$y(\ell) = \alpha \cdot y(\ell - 1) + (1 - \alpha) \cdot x(\ell), \tag{A.1}$$

wobei $x(\ell)$ und $y(\ell)$ jeweils das zeitdiskrete Eingangssignal und das zeitdiskrete Ausgangssignal sind. Mit Hilfe der z-Transformation lässt sich die Impulsantwort $h(\ell)$ dieses zeitdiskreten IIR-Filters bestimmen

$$h(\ell) = (1 - \alpha) \cdot \alpha^{\ell} \cdot u(\ell), \tag{A.2}$$

wobei $u(\ell)$ die zeitdiskrete Sprungfunktion (auch Heaviside-Funktion genannt) ist, die wie folgt definiert wird

$$u(\ell) = \begin{cases} 1, & \ell \geq 0 \\ 0, & \text{sonst.} \end{cases} \tag{A.3}$$

Wird an den Systemeingang eine gespiegelte Sprungfunktion $u(-\ell)$ multipliziert mit der Amplitude a_0 angelegt, d. h. $x(\ell) = a_0 \cdot u(-\ell)$, lässt sich das Ausgangssignal $y(\ell)$, das als eine Antwort des Systems auf diese Anregung mit $a(\ell)$ bezeichnet wird, über die zeitdiskrete Faltung berechnen:

$$\begin{aligned} a(\ell) &= \sum_{k=-\infty}^{\infty} h(k) \cdot x(\ell - k) = a_0 \cdot (1-\alpha) \cdot \sum_{k=-\infty}^{\infty} \alpha^k \cdot u(k) \cdot u(k-\ell) \\ &= a_0 \cdot \alpha^\ell \cdot u(\ell) + a_0 \cdot u(-\ell - 1) = \begin{cases} a_0 \cdot \alpha^\ell, & \ell \geq 0 \\ a_0, & \text{sonst.} \end{cases} \end{aligned} \tag{A.4}$$

Das Eingangssignal $x(\ell) = a_0 \cdot u(-\ell)$ geht mit dem Szenario einher, wenn ein Ton, der eine lange Zeit anwesend war, plötzlich aufhört. Dieses Szenario erlaubt die für diese Arbeit interessante Kenngrößen des IIR-Filters aus (A.1) zu berechnen.

Parametrisierung über die Zeitkonstante[1]: Eine dieser Kenngrößen ist die Zeitkonstante $\tau \in \mathbb{R}_{>0}$, die eine Zeitdauer angibt, die das Ausgangssignal $a(\ell)$ benötigt, um von seinem Startwert $a(0) = a_0$ auf den Wert $a(\tau) = \frac{a_0}{e}$ abzusinken. Für die Berechnung der Zeitkonstante, die ja reellwertig und nicht ganzzahlig ist, wird die Einhüllende $a(t)$ von $a(\ell)$ aus (A.4) als Funktion der herkömmlichen Zeit t gemessen in Sekunden verwendet. Setzt man den Zusammenhang $t = \ell \cdot \Delta T = \ell \cdot \frac{R}{F_s}$ in (A.4) ein, ergibt sie sich zu

$$a(t) = a_0 \cdot \alpha^{\frac{t}{\Delta T}} = a_0 \cdot \alpha^{\frac{F_s}{R} \cdot t} \quad \text{für} \quad t \geq 0. \tag{A.5}$$

Somit resultiert für die Zeitkonstante der rekursiven Glättung (A.1) gemessen in Sekunden:

$$\tau = \frac{\Delta T}{\ln \frac{1}{\alpha}} = \frac{R}{F_s \cdot \ln \frac{1}{\alpha}}. \tag{A.6}$$

Die Verwendung einer Glättungskonstante $\alpha_{\text{ref}} = 0.9$ in der Gleichung (11) in [GH11] für die rekursive Glättung der Sprachpräsenzwahrscheinlichkeit würde somit bei der Abtastfrequenz der Sprachsignale $F_{s,\text{ref}} = 16\,\text{kHz}$ und beim STFT-Fenstervorschub $R_{\text{ref}} = 2^8$ einer Zeitkonstante $\tau \approx 152\,\text{ms}$ entsprechen (siehe auch Abb. A.1):

$$\alpha_{\text{ref}} = 0.9 \quad \text{mit} \quad F_{s,\text{ref}} = 16\,\text{kHz} \quad \text{und} \quad R_{\text{ref}} = 2^8 \quad \Rightarrow \quad \tau \approx 152\,\text{ms}. \tag{A.7}$$

Für $F_{s,\text{ref}} = 16\,\text{kHz}$ und $R_{\text{ref}} = 2^8$ aus [GH11] ist die Zeitkonstante τ als Funktion von α in Abb. A.1 dargestellt.

Würde man Experimente mit den Sprachsignalen abgetastet mit $F_s = 8\,\text{kHz}$ beim Fenstervorschub von $R = 2^9$ durchführen und wäre die rekursive Filterung mit derselben Zeitkonstante $\tau \approx 152\,\text{ms}$ nötig, müsste man eine angepasste Glättungskonstante α verwenden:

$$\alpha = \alpha_{\text{ref}}^{\frac{F_{s,\text{ref}}}{R_{\text{ref}}} \cdot \frac{R}{F_s}}. \tag{A.8}$$

Für die Referenzangaben aus (A.7) ergibt sich mit (A.8) die angepasste Glättungskonstante $\alpha \approx 0.656$, die in derselben Zeitkonstante resultiert:

$$\alpha = 0.656 \quad \text{mit} \quad F_s = 8\,\text{kHz} \quad \text{und} \quad R = 2^9 \quad \Rightarrow \quad \tau \approx 152\,\text{ms}.$$

[1] Man beachte, dass alternativ zur Zeitkonstante τ gemessen in Sekunden auch die Grenzfrequenz f_c gemessen in Hertz für die Parametrisierung der rekursiven Glättung (A.1) verwendet werden könnte, die sich für $\alpha \in (0\,;1)$ als ein BIBO-stabiles Tiefpassfilter darstellt.

Eine Angabe der Zeitkonstante τ statt des Glättungsparameters α scheint zwar eine sinnvolle Alternative für die Parametrisierung der rekursiven Glättung (A.1) zu sein, jedoch benutzerfreundlich ist sie nicht, denn in der Realisierung eines Schätzverfahrens der Glättungsparameter α und nicht die Zeitkonstante τ verwendet wird. Allerdings als eine zusätzliche Angabe lohnt sie sich schon, denn somit wird der Leser darauf aufmerksam gemacht, dass die Glättungskonstante α allein kein universeller Parameter ist, der in allen Systemen mit beliebigen Werten von F_s und R uneingeschränkt verwendet werden kann. Außerdem drückt die Zeitkonstante τ die Eigenschaften des zeitdiskreten IIR-Filters (A.1), der eigentlich auf der diskreten dimensionslosen Zeitachse ℓ arbeitet, in der herkömmlichen Zeit gemessen in Sekunden aus. Somit lassen sich die Auswirkungen der rekursiven Glättung auf die analogen Signale der realen Welt wie die z. B. Sprachsignale besser erklären.

Jedoch, ein Entwurf solcher IIR-Filter wie die rekursive Glättung (A.1) kann auch auf Problemstellungen basieren, die mit der Vorgabe einer Zeitkonstante noch nicht gelöst werden. So wird in [Mar01] die Glättungskonstante für folgende Problemstellung benötigt.

Problemstellung A.1 *Für den Fall eines plötzlichen Abfalls der spektralen Leistung im Spektrogramm von einem hohen Niveau auf ein niedriges Niveau um z. B.* $40\,\text{dB}$ *berechne die Glättungskonstante* α *so, dass die rekursive Glättung* (A.1) *innerhalb einer Zeitdauer von* $64\,\text{ms}$ *vom hohen Niveau auf das niedrige Niveau absinkt.*

Somit ist hier die Verfolgungsgeschwindichkeit der rekursiven Glättung entscheidend. Bei der Parametrisierung der rekursiven Glättung in [LSTW14, RKRS16] wird eigentlich auch die Verfolgungsgeschwindichkeit verwendet, ohne allerdings sie als solche zu definieren.

Alternative Parametrisierung über die Verfolgungsgeschwindichkeit: Als Motivation für die Einführung einer Verfolgungsgeschwindichkeit kann eine solche Größe wie die Schallabklinggeschwindigkeit (engl. *sound decay rate*) angesehen werden, die sich in der Raumakustik als eine alternative Größe zur Nachhallzeit T_{60} (engl. *reverberation time*) etablierte. Der Definition nach gibt die Schallabklinggeschwindigkeit an, wie schnell ein Schallpegel gemessen in einem Raum in Dezibel (dB) abnimmt, nachdem eine akustische Quelle ausgeschaltet wird [AP88]. Ähnlich wird auch die Verfolgungsgeschwindichkeit definiert:

$$v_F \triangleq \frac{20 \cdot \log_{10} \frac{a_0}{a(t)}}{t}, \tag{A.9}$$

die in dB/s gemessen wird. Die Definition von Dezibel in (A.9) offenbart, dass a_0 und $a(t)$ als Feldgrößen angenommen werden, worauf der tiefgesetzte Buchstabe 'F' bei v_F deutet. Setzt man (A.5) in (A.9) ein, stellt man fest, dass bei der rekursiven Glättung erster Ordnung (A.1) die Verfolgungsgeschwindichkeit unabhängig von der absoluten Zeit ist

$$v_F = \frac{F_s}{R} \cdot 20 \log_{10} \frac{1}{\alpha}. \tag{A.10}$$

Bis auf das Vorzeichen entspricht (A.10) der Formel (9) in [LSTW14], wo die damit berechnete Größe als 'benutzerfreundliche Zeitkonstante' bezeichnet wird. Allerdings ist v_F keine Zeitkonstante, sondern eine Größe, die angibt, welchen Pegelunterschied gemessen in dB die rekursive Glättung (A.1) innerhalb einer Sekunde überbrücken kann. Für $F_{s,\text{ref}} = 16\,\text{kHz}$ und $R_{\text{ref}} = 2^8$ aus [GH11] ist v_F als Funktion von α in Abb. A.1 dargestellt. Für das Referenzbeispiel (A.7) ergibt sich mit (A.10) für die Verfolgungsgeschwindichkeit

$$\alpha_{\text{ref}} = 0.9 \quad \text{mit} \quad F_{s,\text{ref}} = 16\,\text{kHz} \quad \text{und} \quad R_{\text{ref}} = 2^8 \quad \Rightarrow \quad v_F \approx 57.2\,\text{dB/s}. \tag{A.11}$$

Man beachte, dass die Zeitkonstante τ als eine Zeitdauer interpretiert werden kann, nach der bei Glättung einer Feldgröße der Signalpegel um $v_F \cdot \tau = 20 \log_{10} e \approx 8.7\,\text{dB}$ abnimmt. Somit erweist sich die Verfolgungsgeschwindichkeit v_F als eine Kenngröße der rekursiven Glättung, die universeller als die Zeitkonstante τ zu sein scheint. Aus diesem Grund wird die Verfolgungsgeschwindichkeit v_F neben der Angabe der Glättungskonstante α angegeben.

O.B.d.A. kann man ein Startwert-zu-Aktuellwert Verhältnis (engl. *start-to-current value ratio*, SCR) als Funktion der Zeit $\text{SCR}(t)$ definieren, das unter Berücksichtigung dessen, dass a_0 und $a(t)$ Feldgrößen sind, auch in Dezibel angegeben werden kann

$$\text{SCR}(t) \triangleq \frac{a_0}{a(t)}, \qquad \text{(A.12)} \qquad \text{SCR}_{\text{dB}}(t) \triangleq 20 \cdot \log_{10} \text{SCR}(t). \qquad \text{(A.13)}$$

Mit (A.9) ist das $\text{SCR}_{\text{dB}}(t)$ eine lineare Funktion der Zeit $\text{SCR}_{\text{dB}}(t) = v \cdot t$ und kann als eine Pegelveränderung gemessen in dB betrachtet werden, die innerhalb einer bestimmten Zeit vom Ausgangssignal der rekursiven Glättung zurückgelegt wird. Möchte man die rekursive Glättung so parametrisieren, dass sie fähig ist, einen bestimmten absoluten SCR-Wert innerhalb der vorgegebenen Zeit t_{SCR} zu überbrücken, muss man unter Verwendung von (A.10) folgende Glättungskonstante wählen:

$$\alpha = \text{SCR}^{-\frac{R}{F_s \cdot t_{\text{SCR}}}}. \qquad \text{(A.14)}$$

Wird der zu überbrückende Pegelunterschied in Dezibel vorgegebenen, wird die Glättungskonstante wie folgt berechnet

$$\alpha_{\text{SCR}} = 10^{-\frac{R \cdot \text{SCR}_{\text{dB}}}{20 \cdot F_s \cdot t_{\text{SCR}}}}. \qquad \text{(A.15)}$$

(A.14) und (A.15) könnten bei der Lösung der **Problemstellung A.1** helfen, wenn dort eine Feldgröße gefiltert werden würde, was allerdings nicht der Fall ist. Und da bei den Leistungsgrößen ein Dezibel, der ja direkt in die Definition der Verfolgungsgeschwindichkeit (A.9) einfließt, etwas anders als bei den Feldgrößen definiert wird, muss die rekursive Glättung von Leistungsgrößen gesondert betrachtet werden.

Rekursive Glättung von Leistungsgrößen: Da in [Mar01] das hohe Niveau mit der mittleren spektralen Leistung des ungestörten Sprachsignals $a_0 \to \mathbb{E}\left[|S(k,\ell)|^2\right]$ und das niedrige mit der des Störsignals $a(t) \to \mathbb{E}\left[|D(k,\ell)|^2\right]$ einhergehen, kann in diesem Fall das Verhältnis $\frac{a_0}{a(t)}$ o.B.d.A. als SNR bezeichnet werden[2]. Unter Berücksichtigung dessen, dass ein Dezibel bei den Leistungsgrößen etwas anders deklariert wird, gelten folgende Definitionen in diesem Fall

$$\text{SNR}(t) \triangleq \frac{a_0}{a(t)}, \qquad \text{(A.16)} \qquad \text{SNR}_{\text{dB}}(t) \triangleq 10 \cdot \log_{10} \text{SNR}(t). \qquad \text{(A.17)}$$

Mit (A.17) kann die Verfolgungsgeschwindichkeit bei Filterung von Leistungsgrößen (engl. *power quantity*) etwas anders als in (A.9) definiert werden:

$$v_P \triangleq \frac{\text{SNR}_{\text{dB}}(t)}{t}. \qquad \text{(A.18)}$$

Setzt man (A.5), (A.16) und (A.17) in (A.9) ein, resultiert die Verfolgungsgeschwindichkeit

$$v_P = \frac{F_s}{R} \cdot 10 \log_{10} \frac{1}{\alpha}, \qquad \text{(A.19)}$$

[2]Laut Formulierung in [Mar01] ist das Verhältnis $\frac{a_0}{a(t)}$ eigentlich das *a priori* SNR $\frac{\mathbb{E}\left[|S(k,\ell)|^2\right]}{\mathbb{E}\left[|D(k,\ell)|^2\right]}$, wie in (2.22).

die ähnlich wie v_F unabhängig von der absoluten Zeit ist. Man beachte, dass bei den gleichen Parametern F_s, R und α die Verfolgungsgeschwindichkeit v_P um den Faktor 2 kleiner ist als die Verfolgungsgeschwindichkeit $v_P = v_F/2$. Für $F_{s,\text{ref}} = 16\,\text{kHz}$ und $R_{\text{ref}} = 2^8$ aus [GH11] ist die Verfolgungsgeschwindigkeit v_P als Funktion von α in Abb. A.1 dargestellt.

Soll die rekursive Glättung (A.1), in der das Eingangssignal $x(\ell)$ eine Leistungsgröße ist, innerhalb einer Zeitdauer von t_{SNR} die Pegelveränderung SNR oder SNR_{dB} erreichen, soll der Glättungsparameter wie folgt gewählt werden:

$$\alpha = \text{SNR}^{-\frac{R}{F_s \cdot t_{\text{SNR}}}}, \qquad \text{(A.20)} \qquad\qquad \alpha_{\text{SNR}} = 10^{-\frac{R \cdot \text{SNR}_{\text{dB}}}{10 \cdot F_s \cdot t_{\text{SNR}}}}. \qquad \text{(A.21)}$$

Die Gleichung (A.20) ist dabei die Gleichung (12) in [Mar01], wo ein absoluter SNR-Wert (nicht in dB gemessen) eingesetzt wird. Erwartungsgemäß sind die Gleichungen (A.14) und (A.20) sehr identisch, denn es geht in den beiden Fällen um die absolute Pegelveränderung der gefilterten Größe. Müsste also die rekursive Glättung (A.1) in der **Problemstellung A.1** innerhalb der Zeitdauer von $t_{\text{SNR}} = 64\,\text{ms}$ nach der Pegelveränderung den Pegelunterschied von z. B. $\text{SNR}_{\text{dB}} = 40\,\text{dB}$ zurücklegen können, soll die Glättungskonstante bei den in (A.7) gegebenen Werten von $F_{s,ref} = 8\,\text{kHz}$ und $R_{\text{ref}} = 2^7$ laut (A.21) zu $\alpha = 0.1$ gewählt werden.

Man beachte, für den Fall $\text{SCR}_{\text{dB}} = \text{SNR}_{\text{dB}}$ und $t_{\text{SCR}} = t_{\text{SNR}}$ bei gleichen Parametern F_s und R gilt mit (A.15) und (A.21) der Zusammenhang $\alpha_{\text{SCR}} = \sqrt{\alpha_{\text{SNR}}}$. Der Unterschied zwischen α_{SCR} und α_{SNR} ist der Grund, warum SCR neben SNR gesondert eingeführt wird. Hätte man diesen Unterschied nicht beachtet, würde die Angabe der Verfolgungsgeschwindichkeit immer in Dezibel erfolgen, wie sie bei den Feldgrößen definiert werden. Dies könnte dann bei der Filterung von Leistungsgrößen, wie dies in [LSTW14] oder in [RKRS16] der Fall ist, für vermeidbare Missverständnisse sorgen. Die Notwendigkeit der Unterscheidung, ob eine Feldgröße oder eine Leistungsgröße gefiltert werden, kann einerseits als ein Nachteil empfunden werden. Andererseits sorgt sie für mehr Klarheit in der Hinsicht, was von der rekursiven Filterung bewirkt wird.

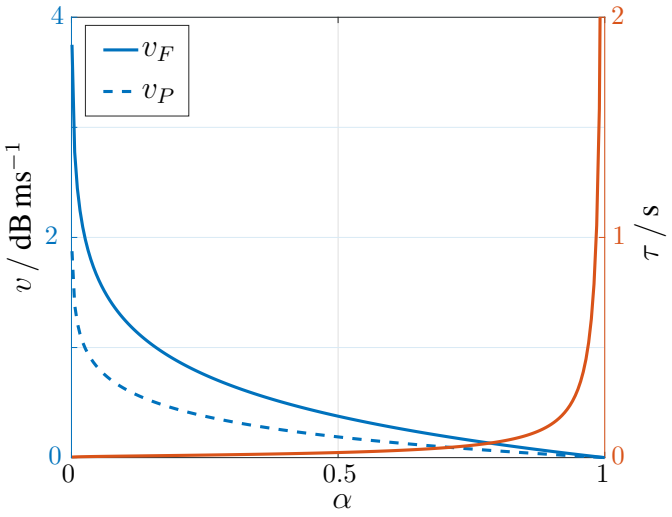


Abbildung A.1.: Die Zeitkonstante τ und die Verfolgungsgeschwindigkeiten v_F und v_P als Funktionen der Glättungskonstante α für die Abtastfrequenz $F_{s,\text{ref}} = 16\,\text{kHz}$ und für den Rahmenvorschub $R_{\text{ref}} = 2^8$ aus [GH11].

Akronyme

AMS ..... ***A**nalyse-**M**odifikation-**S**ynthese*
ANSI ..... ***A**merican **N**ational **S**tandards **I**nstitute*
ASR ..... ***A**utomatic **S**peech **R**ecognition*
BAN ..... ***B**lind **A**nalytical **N**ormalization*
BCE ..... ***B**inary **C**ross **E**ntropie*
BIBO ..... ***B**ounded **I**put **B**ounded **O**utput*
BLSTM ..... ***B**idirectional **L**ong **S**hort-**T**erm **M**emory*
BM ..... ***B**iaskompensation und **M**inimumsuche*
BS ..... ***B**ayes-motivated **S**moothing*
BSMS ..... ***B**ayes-motivated **S**moothed **M**inimum **S**tatistics*
DD ..... ***D**ecision **D**irected*
DFT ..... ***D**iscrete **F**ourier **T**ransform*
DNN ..... ***D**eep **N**eural **N**etwork*
ELU ..... ***E**xponential **L**inear **U**nit*
EM ..... ***E**rror **M**onitoring*
EMCRA ..... ***E**nhanced **M**inima **C**ontrolled **R**ecursiv **A**veraging*
FFT ..... ***F**ast **F**ourier **T**ransform*
FIR ..... ***F**inite **I**mpuls **R**esponse*
FP ..... ***F**ixed **P**riors*
GDD ..... ***G**eneralized **D**ecision **D**irected*
GEV ..... ***G**eneralized **E**igen**v**alue*
GGV ..... ***G**eneralisierte **G**amma **V**erteilung*
GLR ..... ***G**eneralized **L**ikelihood **R**atio*
GPU ..... ***G**raphics **P**rocessing **U**nits*
GSA ..... ***G**eneralisierte **S**pektrale **A**mplitude*
GSS ..... ***G**eneralisierte **S**pektrale **S**ubtraktion*
GW ..... ***G**eneralized **W**eighted*
HMM ..... ***H**idden **M**arkov **M**odels*
IBM ..... ***I**deal **B**inary **M**ask*
ICASSP ..... ***I**nternational **C**onference on **A**coustics **S**peech and **S**ignal **P**rocessing*
IIR ..... ***I**nfinite **I**mpuls **R**esponse*
IMCRA ..... ***I**mproved **M**inima **C**ontrolled **R**ecursiv **A**veraging*
IN ..... ***In**put*
IRM ..... ***I**deal **R**atio **M**ask*
ISD ..... ***I**takuro-**S**aito **D**istanz*
ISTFT ..... ***I**nverse **S**hort-**T**ime **F**ourier **T**ransform*
ITG ..... ***I**nformations**t**echnische **G**esellschaft*

ITU ..... ***I**nternational **T**elecommunication **U**nion*
KM ..... ***K**ombiniertes **M**aß*
KN ..... ***K**onsistente **N**ormalverteilung*
LDS ..... ***L**eistungs**d**ichte**s**pektrum*
LEM ..... ***L**og-**E**rror **M**ean*
LEV ..... ***L**og-**E**rror **V**ariance*
LGSA ..... ***L**ogaritthmische **G**eneralisierte **S**pektrale **A**mplitude*
LSA ..... ***L**og-**S**pectral **A**mplitude*
LSTM ..... ***L**ong **S**hort-**T**erm **M**emory*
MAP ..... ***M**aximum **A**-**P**osteriori*
MAPB ..... ***M**aximum **A**-**P**osteriori* **b**asiert
MCRA ..... ***M**inima **C**ontrolled **R**ecursiv **A**veraging*
ML ..... ***M**aximum **L**ikelihood*
MLEM ..... ***M**inimum **L**og-**E**rror **M**ean*
MLEV ..... ***M**inimum **L**og-**E**rror **V**ariance*
MMSE ..... ***M**inimum **M**ean **S**quared **E**rror*
MOS ..... ***M**ean **O**pinion **S**core*
MOS-LQO ..... ***M**ean **O**pinion **S**core - **L**istening **Q**uality **O**bjective*
MS ..... ***M**inimum **S**tatistics*
MSE ..... ***M**ean **S**quared **E**rror*
MSK ..... ***M**el-spekltale **K**oeffizienten*
NB ..... ***N**arrow-**B**and*
NPP ..... ***N**oise-only **P**resence **P**robability*
NWNR ..... ***N**ichtstationäres **W**eißes **N**ormalverteiltes **R**auschen*
OLA ..... ***O**ver**l**ap-**A**dd*
OMLSA ..... ***O**ptimally **M**odified **L**og-**S**pectral **A**mplitude*
OS ..... ***O**ptimally **S**moothed*
OSMS ..... ***O**ptimally **S**moothed **M**inimum **S**tatistics*
OUT ..... ***Out**put*
PESQ ..... ***P**erceptual **E**valuation of **S**peech **Q**uality*
PGSS ..... ***P**arametrische **G**eneralisierte **S**pektrale **S**ubtraktion*
PGSSR ..... *PGSS mit **R**andbedingung*
PSD ..... ***P**ower **S**pectral **D**ensity*
RA ..... ***R**ecursive **A**veraging*
ReLU ..... ***Re**ctified **L**inear **U**nit*
RLDS ..... ***R**auschl**e**istungs**d**ichte**s**pektrum*
RMSE ..... ***R**oot **M**ean **S**quared **E**rror*
RNN ..... ***R**ekurrentes **N**euronales **N**etz*
ROC ..... ***R**eceiver **O**perating **C**haracteristic*
RSG ..... ***R**esearch **S**tudy **G**roup*
SA ..... ***S**pektrale **A**mplitude*
SAP ..... ***S**eech **A**bsence **P**robability*
SCR ..... ***S**tart-to-**C**urrent value **R**atio*
SDM ..... ***S**pectral **D**istance **M**easure*
SL ..... ***S**pektrale **L**eistungssubtraktion*
SNR ..... ***S**ignal-to-**N**oise **R**atio*

SNT ..... ***S**ubspace **N**oise **T**racking*
SPIB ..... ***S**ignal **P**rocessing **I**nformation **B**ase*
SPP ..... ***S**eech **P**resence **P**robability*
STFT ..... ***S**hort-**T**ime **F**ourier **T**ransform*
STM ..... ***S**hort-**T**ime **M**odulation*
STOI ..... ***S**hort-**T**ime **O**bjective **I**ntelligibility*
SuGAR ..... ***Su**per **G**aussian **A**mplitude **R**oot*
SWNR ..... ***S**inusförmiges **W**eißes **N**ormalverteiltes **R**auschen*
TIMIT ..... ***T**exas **I**nstrument and **M**assachusetts **I**nstitute of **T**echnology*
VAD ..... ***V**oice **A**ktivity **D**etection*
VDF ..... ***V**erteilungs**d**ichte**f**unktion*
WB ..... ***W**ide-**B**and*
WNR ..... ***W**eißes **N**ormalverteiltes **R**auschen*
WNZ ..... ***W**eißer **N**ormalverteilter **Z**ufallsprozess*
WSJ0 ..... ***W**all **S**treet **J**ournal*
ZF ..... ***Z**eit-**F**requenz*
CHiME ..... ***C**omputational **H**earing **i**n **M**ultisource **E**nvironments*

Formelzeichen

Allgemeine Notation

- Schätzgrößen werden durch ein Dach-Akzent über einem Symbol gekennzeichnet wie z. B. $\hat{s}(n)$ oder $\hat{\theta}$.

Römische Formelzeichen

a_R	Verschiebungsfaktor der STFT/ISTFT
$a(\ell)$	Antwort eines zeitdiskreten Filters auf eine Sprungfunktion
b^{MAPB}	Bias des MAPB-Postprozessors
c_i, c_{min}	Parameter der Approximation $c_{A,\text{var}}(\rho)$
c_β	Konstante in der vorgeschlagenen LGSA-Filterfunktion
c_ρ	Konstante im vorgechlagenen SPP-Schätzer
c_{var}	Korrekturfaktor der Varianzen var$(\bar{\gamma}_\rho \vert H_0)$ und var$(\bar{\gamma}_\rho \vert H_1)$
$d(n)$	Zeitdiskretes Störsignal
$d_{\text{NS}}(n)$	Instationaritätsfaktor (*degree of nonstationarity*)
$D(k,\ell)$	STFT Koeffizient eines Störsignals $d(n)$ im (k,ℓ)-ten ZF-Punkt
$D_\beta(k,\ell)$	Generalisierte spektrale Amplitude von $D(k,\ell)$
D_{MS}	Anzahl der STFT-Rahmen im MS-Fenster für Minimumsuche
e	Eulerische Zahl $2.7182...$
E_w	Parameter der verwendeten Gewichte $w(k_\Delta, \ell_\Delta)$
f_c	Grenzfrequenz eines Tiefpassfilters
F_s	Abtastrate eines zeitdiskreten Signals
$G(k,\ell)$	Spektrale Filterfunktion im (k,ℓ)-ten ZF-Punkt
G_{Art}	Spektrale Filterfunktion bestimmter Art (z. B. Wieder-Filter G_{WF})
G_{H_0}	Spektrale Filterfunktion bei Sprachsignalabwesenheit
$G_{H_1}(k,\ell)$	Spektrale Filterfunktion bei Sprachsignalpräsenz
G_{max}	Maximaler Wert einer spektralen Filterfunktion
G_{min}	Minimaler Wert einer spektralen Filterfunktion
G_{ref}	Referenzwert einer spektralen Filterfunktion für Angabe in Dezibel
$h(\ell)$	Impulsantwort eines zeitdiskreten Filters
$H(k,\ell)$	Binäre Zufallsvariable für Sprachsignalpräsenz/-abwesenheit
H_0	Hypothese der Sprachsignalabwesenheit
H_1	Hypothese der Sprachsignalpräsenz
$I(k,\ell)$	Binärer Indikator der Sprachsignalpräsenz im (k,ℓ)-ten ZF-Punkt
j	Imaginäre Zahl
k	Ganzzahliger STFT-Frequenzindex

k_{Nyq} ..... Ganzzahliger Frequenzindex der Nyquist-Frequenz
k_Δ ..... Ganzzahliger Frequenzindex in einer Zeit-Frequenz-Umgebung
Δk ..... Einseitige Anzahl benachbarter Frequenzbänder in einer Zeit-Frequenz-Umgebung
K ..... DFT-Länge und Anzahl der Abtastwerte in einem STFT-Block
ℓ ..... Ganzzahliger STFT-Blockindex
ℓ_Δ ..... Ganzzahliger Blockindex einer Zeit-Frequenz-Umgebung
$\Delta\ell$ ..... Anzahl vergangener Rahmen in einer ZF-Umgebung
L ..... Anzahl der STFT-Rahmen in einem Signal
$M_D(k,\ell)$ ..... Spektrale Maske des Störsignals $d(n)$ im ZF-Bereich
n ..... Ganzzahliger Zeitindex
N ..... Gesamte Anzahl der ZF-Punkte in einer ZF-Umgebung
N_{IN} ..... Anzahl der Eingangsknoten einer DNN-Schicht
N_{OUT} ..... Anzahl der Ausgangsknoten einer DNN-Schicht
$\mathcal{P}(k,\ell)$ ..... *A posteriori* Sprachpräsenzwahrscheinlichkeit
$\mathcal{P}_{\text{dropout}}$ ..... Dropout-Wahrscheinlichkeit
$\mathcal{P}_{\text{frame}}(\ell)$ ..... Sprachpräsenzwahrscheinlichkeit des ℓ-ten STFT-Rahmens
$\mathcal{P}_{\text{glob}}(k,\ell)/\mathcal{P}_{\text{loc}}(k,\ell)$. *A posteriori* SPP im (k,ℓ)-ten ZF-Punkt ausgerechnet auf den Beobachtungen einer globalen / lokalen Zeit-Frequenz-Umgebung
P_F ..... Falschalarmwahrscheinlichkeit
P_M ..... Wahrscheinlichkeit des verpassten Treffers (*missed-hit probability*)
$q(k,\ell)$ ..... *A priori* Sprachpräsenzwahrscheinlichkeit im (k,ℓ)-ten ZF-Punkt
$r_{\text{glob}}/r_{\text{loc}}$ ..... Korrelationskoeffizient einer globalen / lokalen Umgebung
R ..... STFT/ISTFT Rahmenvorschub
$s(n)$ ..... Zeitdiskretes ungestörtes Sprachsignal
$\hat{s}(n)$ ..... Schätzung eines ungestörten Sprachsignals
$S(k,\ell)$ ..... STFT Koeffizient eines ungestörten Sprachsignals $s(n)$
$S_\beta(k,\ell)$ ..... Generalisierte spektrale Amplitude von $S(k,\ell)$
t ..... Zeit gemessen in Sekunden
ΔT ..... Zeitlicher Abstand zwischen den aufeinander folgenden STFT-Rahmen gemessen in Sekunden
$u(\ell)$ ..... Zeitdiskrete Einheitssprungfunktion (auch Heaviside-Funktion)
$U(x)$ ..... Einheitssprungfunktion (engl. *unit step function*)
U_{MS} ..... Anzahl der Unterfenster in der MS-Minimumsuche
v_F/v_P ..... Verfolgungsgeschwindigkeit einer rekursiven Glättung einer Feldgröße / einer Leistungsgröße
V_{MS} ..... Anzahl der STFT-Rahmen im Unterfenster der MS-Minimumsuche
$w(k_\Delta,\ell_\Delta)$ ..... Gewichte der Beobachtungen einer ZF-Umgebung
$y(n)$ ..... Gestörtes Sprachsignal im Zeitbereich
$Y(k,\ell)$ ..... STFT Koeffizient eines gestörten Sprachsignals $y(n)$
$Y_\beta(k,\ell)$ ..... Generalisierte spektrale Amplitude von $Y(k,\ell)$

Griechische Formelzeichen

α ..... Glättungskonstante einer rekursiven Glättung erster Ordnung
α_ρ ..... Gewichtsfaktor des generalisierten *Decision-Directed* Verfahrens

α_{DD} ..... Gewichtsfaktor des *Decision-Directed* Verfahrens
α_{max} ..... Maximaler Wert einer Glättungskonstante
α_{min} ..... Minimaler Wert einer Glättungskonstante
β ..... Kompressionsfaktor spektraler Amplituden
β_{max} ..... Maximaler Biaskompensationsfaktor des MAPB-Postprozessors
$\gamma(k,\ell)$ ..... *A posteriori* SNR im (k,ℓ)-ten ZF-Punkt
$\bar{\gamma}$ ..... Geglättete oder gemittelte *a posteriori* SNR
$\gamma_\rho(k,\ell)$ ..... Generalisiertes *a posteriori* SNR im (k,ℓ)-ten ZF-Punkt
$\bar{\gamma}_\rho$ ..... Gewichtetes generalisiertes *a posteriori* SNR
$\Delta(k,\ell)$ ..... Logarithmischer Schätzfehler im (k,ℓ)-ten ZF-Punkt
η ..... Zweiter Formparameter einer generalisierten Gamma-Verteilung
$\hat{\theta}_\rho^{\text{DD}}(k,\ell)$ ..... Generalisiertes *a priori* SNR im (k,ℓ)-ten ZF-Punkt geschätzt mit dem *Decision-Directed* Verfahren
ϑ ..... Formparameter einer Weibull-Verteilung
ϑ_D ..... Entscheidungsschwelle für Berechnung von $\text{IBM}_D(k,\ell)$
$\lambda_D(k,\ell)$ ..... Leistungsdichtespektrum eines Störsignals $d(n)$ oder auch Rauschleistungsdichtespektrum
$\lambda_S(k,\ell)$ ..... Leistungsdichtespektrum eines ungestörten Sprachsignals $s(n)$
$\lambda_Y(k,\ell)$ ..... Leistungsdichtespektrum eines gestörten Sprachsignals $y(n)$
$\Lambda(k,\ell)$ ..... Generalisiertes Likelihood-Verhältnis im (k,ℓ)-ten ZF-Punkt
Λ_{min} ..... Minimaler Wert des generalisierten Likelihood-Verhältnisses
μ_X ..... Mittelwert einer Zufallsvariablen X
ν ..... Freiheitsgrad verschiedener Chi-Quadrat Verteilungen
ν_0 ..... Konstanter Freiheitsgrad des MAPB-Postprozessors
$\Delta\nu$ ..... Parameter der alternativen Steuerungsfunktion des OSMS-Verfahrens in (5.21) oder Regelbereich der Bandbreitenanpassung des MAPB-Postprozessors in (7.34)
$\breve{\zeta}(k,\ell)$ ..... Geglätteter *a priori* SNR-Schätzwert im (k,ℓ)-ten ZF-Punkt
ζ_ρ ..... Momentanes generalisiertes *a priori* SNR
$\tilde{\zeta}_\rho$ ..... Generalisierter propagierter *a priori* SNR-Schätzwert
$\xi(k,\ell)$ ..... *A priori* SNR im (k,ℓ)-ten ZF-Punkt
ξ_i ..... Parameter der Funktion $\bar{\xi}_{\text{fix}}(\rho)$
$\xi_\rho(k,\ell)$ ..... Generalisiertes *a priori* SNR im (k,ℓ)-ten ZF-Punkt
ξ_{min} ..... Minimaler Wert des *a priori* SNR
ξ_{ref} ..... Referenzwert des *a priori* SNR für Angabe in Dezibel
$\tilde{\xi}$ ..... Propagierter *a priori* SNR-Schätzwert
$\bar{\xi}_{\text{fix}}$ ..... Fixiertes *a priori* SNR
π ..... Kreiszahl $3.1415926...$
ρ ..... Exponent oder Kompressionsfaktor generalisierter SNR-Größen
ρ_{fix} ..... Fixierter Exponent generalisierter spektraler SNR-Größen
σ_X^2 ..... Varianz einer Zufallsvariablen X
σ^{MAPB} ..... Empirische Standardabweichung des MAPB-Postprozessors
τ ..... Zeitkonstante eines Filters
τ^2 ..... Skalierungsparameter inverser skalierter Chi-Quadrat Verteilung
$\Upsilon(\ell)$ ..... Eingangsseitiges globales SNR_{IN} im ℓ-ten STFT-Rahmen in dB

$\Upsilon_0(\ell)$	Absoluter Wert von $\Upsilon(\ell)$
$\varphi_S(k,\ell)/\varphi_Y(k,\ell)$	Phasenkomponente eines STFT-Koeffizienten $S(k,\ell)/Y(k,\ell)$
ω_0	Initiale Gewichte eines DNNs
$\omega_\mathrm{a}(n)$	Analyse-Fenster einer STFT
$\omega_\mathrm{s}(n)$	Synthese-Fenster einer ISTFT

Spezielle Symbole, Operatoren, Funktionen und Verteilungsdichtefunktionen

$\mathbb{C}$	Menge komplexer Zahlen
$\mathbb{N}$	Menge aller natürlichen Zahlen $\{0, 1, 2, \ldots\}$
$\mathbb{N}_{>0}$	Menge natürlicher Zahlen ohne Null $\{1, 2, 3 \ldots\}$
$\mathbb{N}_{>1}$	Menge natürlicher Zahlen größer Eins $\{2, 3, 4 \ldots\}$
$\mathbb{R}$	Menge aller reellen Zahlen
$\mathbb{R}_{>0}$	Menge positiver reeller Zahlen ohne Null
$\mathrm{cov}(\cdot)$	Operator für Berechnung einer Kovarianz
$\mathbb{E}[\cdot]$	Erwartungswert-Operator
$\max(\cdot)$	Maximum-Operator
$\min(\cdot)$	Minimum-Operator
$\Pr(\cdot)$	Wahrscheinlichkeitsoperator
$\mathrm{var}(\cdot)$	Operator für Berechnung einer Varianz
$\arg\max_x f(x)$	Operator zum Finden einer Maximumstelle einer Funktion $f(x)$
$\Gamma(x)$	Gamma-Funktion
$M(x,y,z)$	Konfluente hypergeometrische Funktion
$\Phi(x)$	Verteilungsfunktion einer Standardnormalverteilung
$\lvert z\rvert$	Betrag einer komplexen Zahl z
$\angle z$	Phase einer komplexen Zahl z
$\chi^2(x;\, \nu, \lambda_X)$	Skalierte Chi-Quadrat-Verteilung einer Zufallsvariablen X mit Freiheitsgrad ν und mit Skalierungsparameter λ_X
$\mathrm{Exp}(x; \lambda_X)$	Exponentialverteilung einer Zufallsvariablen X mit Mittelwert λ_X
$\Gamma(\lambda; \mu_\lambda, \sigma_\lambda^2)$	Gamma-Verteilung einer Zufallsvariablen mit Realisierung λ, Mittelwert μ_λ und Varianz σ_λ^2
$\mathrm{GenGam}(x; \lambda_X, \beta, \eta)$	Generalisierte Gamma-Verteilung einer Zufallsvariablen X mit Skalierungsparameter λ und zwei Formparametern β und η
Inv-$\chi^2(x;\, \nu, \tau_X^2)$	Inverse skalierte Chi-Quadrat-Verteilung einer Zufallsvariablen X mit Freiheitsgrad ν und Skalierungsparameter τ_X^2
$\mathcal{N}(x;\, \mu_X, \sigma_X^2)$	Normalverteilung einer reellwertigen Zufallsvariablen X mit Mittelwert μ_X und Varianz σ_X^2
$\mathcal{N}_\mathbb{C}(x;\, \lambda_X)$	Mittelwertfreie Normalverteilung einer komplexwertigen Zufallsvariablen X mit Varianz λ_X
$p_X(x)$	Verteilungsdichtefunktion einer Zufallsvariablen X mit Realisierungen x
$\mathrm{Weib}(x;\, \lambda_X,\, \beta)$	Weibull-Verteilung einer reellwertigen Zufallsvariablen X mit Skalierungsparameter λ_X und Formparameter β

Abbildungsverzeichnis

Tabellenverzeichnis

Literatur

[AC07] 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, May 2007.

[ADHG04] C. Avendano, L. Deng, H. Hermansky und B. Gold: „The analysis and representation of speech“, *Speech Processing in the Auditory System*, S. 231–308, Springer, 2004.

[AP88] H. Arau-Puchades: „An improved reverberation formula“, *Acta Acustica united with Acustica*, Band 65(4), S. 163–180, May 1988.

[Ars06] L. M. Arslan: „Modified Wiener filtering“, *Signal Processing*, Band 86(2), S. 267–272, July 2006.

[Ast10] R. F. Astudillo: *Integration of short-time fourier domain speech enhancement and observation uncertainty techniques for robust automatic speech recognition*, Ph.D. thesis, Technical University Berlin, Feb. 2010.

[AW06] I. Andrianakis und P. R. White: „MMSE speech spectral amplitude estimators with chi and gamma speech priors“, *In Proc. of IEEE Int. Conf. on Acoustics Speech and Signal Processing*, Band 3, S. 1068–1071, May 2006.

[AW09] I. Andrianakis und P. R. White: „Speech spectral amplitude estimators using optimally shaped Gamma and Chi priors“, *Speech Communication*, Band 51(1), S. 1–14, Jan. 2009.

[BA11] B. J. Borgström und A. Alwan: „Log-spectral amplitude estimation with generalized Gamma distributions for speech enhancement“, *In Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, S. 4756–4759, May 2011.

[BBB+13] I. Baumgarten, S. Brathuhn, D. Bürgi, M. Müller und P. Rechenberg-Winter: „Warum das laute Lesen zum stillen Schreiben gehört: Schreibwerkstätten als Ort der Selbstvergewisserung“, *Leitfaden*, Band 2(3), S. 85–90, Aug. 2013.

[BCH11a] J. Benesty, J. Chen und E. A. Habets: *Speech enhancement in the STFT domain*, Springer Science & Business Media, 2011.

[BCH11b] J. Benesty, J. Chen und Y. Huang: „Speech enhancement in the Karhunen-Loève expansion domain“, *Synthesis Lectures on Speech and Audio Processing*, Band 7(1), S. 1–112, Jan. 2011.

[BCR75] L.-J. Boë, M. Contini und H. Rakotofiringa: „Étude statistique de la fréquence laryngienne“, *Phonetica*, Band 32(1), S. 1–23, 1975.

[BGM08] C. Breithaupt, T. Gerkmann und R. Martin: „A novel a priori SNR estimation approach based on selective cepstro-temporal smoothing“, *In Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, S. 4897–4900, Apr. 2008.

[Bib84] Bibel: *Altes und Neues Testament (Bibeltext in der revidierten Fassung nach der Übersetzung Martin Luthers)*, Deutsche Bibelgesellschaft, 1984.

[Bis06] C. M. Bishop: *Pattern Recognition and Machine Learning (Information Science and Statistics)*, Springer-Verlag New York, Secaucus, USA, 2006.

[BKM08] C. Breithaupt, M. Krawczyk und R. Martin: „Parameterized MMSE spectral magnitude estimation for the enhancement of noisy speech“, *In Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, S. 4037–4040, Apr. 2008.

[Bla98] K. Blankenburg: „Correct usage of quantities, units and equations (Ill)“, *NEWS-ROHDE AND SCHWARZ*, S. 26–27, 1998.

[BM10] C. Breithaupt und R. Martin: „Analysis of the decision-directed SNR estimator for speech enhancement with respect to low-SNR and transient conditions“, *IEEE Transactions on Audio, Speech, and Language Processing*, Band 19(2), S. 277–289, Apr. 2010.

[BMC05] J. Benesty, S. Makino und J. Chen: *Speech Enhancement*, Springer Science & Business Media, 2005.

[BMVW15] J. Barker, R. Marxer, E. Vincent und S. Watanabe: „The third "CHiME" speech separation and recognition challenge: Dataset, task and baselines“, *In Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, S. 504–511, Dec. 2015.

[BMVW16] J. Barker, R. Marxer, E. Vincent und S. Watanabe: „The third CHiME speech separation and recognition challenge: Analysis and outcomes“, *Computer Speech and Language*, Dec. 2016.

[Bol79] S. Boll: „Suppression of acoustic noise in speech using spectral subtraction“, *In Proc. of IEEE Transactions on Acoustics, Speech and Signal Processing*, Band 27(2), S. 113–120, Apr. 1979.

[Box79] G. E. Box: „Robustness in the strategy of scientific model building“, *Robustness in Statistics*, Band 1, S. 201–236, May 1979.

[Bri74] D. R. Brillinger: „Fourier analysis of stationary processes“, *In Proc. of the IEEE*, Band 62(12), S. 1628–1643, Dec. 1974.

[Bri01] D. R. Brillinger: *Time series: Data Analysis and Theory*, Band 36, Siam, 2001.

[Bro09] P. Brown: „Speech enhancement: WO Patent App. PCT/US2008/010,591“, Mar. 2009.

[BSM79] M. Berouti, R. Schwartz und J. Makhoul: „Enhancement of speech corrupted by acoustic noise“, *In Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Apr. 1979.

[Byr94] D. Byrd: „Relations of sex and dialect to reduction“, *Speech Communication*, Band 15(1-2), S. 39–54, Oct. 1994.

[Cap94] O. Cappe: „Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor“, *IEEE Transactions on Speech and Audio Processing*, Band 2(2), S. 345–349, Apr. 1994.

[CB01a] I. Cohen und B. Berdugo: „Spectral enhancement by tracking speech presence probability in subbands“, *In Proc. of Int. Workshop on Hands-Free Speech Communication*, S. 95–98, Apr. 2001.

[CB01b] I. Cohen und B. Berdugo: „Speech enhancement for non-stationary noise environments“, *Signal Processing*, Band 81(11), S. 2403–2418, Nov. 2001.

[CB02] I. Cohen und B. Berdugo: „Noise Estimation by Minima Controlled Recursive Averaging for Robust Speech Enhancement“, *IEEE Signal Processing Letters*, Band 9(1), S. 12–15, Jan. 2002.

[CBHD06] J. Chen, J. Benesty, Y. Huang und S. Doclo: „New insights into the noise reduction Wiener filter“, *IEEE Transactions on Audio, Speech, and Language Processing*, Band 14(4), S. 1218–1234, 2006.

[CBHD08] J. Chen, J. Benesty, Y. A. Huang und E. J. Diethorn: „Fundamentals of noise reduction“, *Springer Handbook of Speech Processing*, S. 843–872, Springer, 2008.

[CG08] I. Cohen und S. Gannot: „Spectral enhancement methods“, *Springer Handbook of Speech Processing*, S. 873–902, Springer Berlin Heidelberg, 2008.

[CHDHU16] A. Chinaev, J. Heymann, L. Drude und R. Haeb-Umbach: „Noise-Presence-Probability-Based Noise PSD Estimation by Using DNNs“, *In Proc. of the 12th ITG Symposium on Speech Communication*, S. 26–30, Oct. 2016.

[CHHU16] A. Chinaev, J. Heitkaemper und R. Haeb-Umbach: „A Priori SNR Estimation Using Weibull Mixture Model“, *In Proc. of the 12th ITG Symposium on Speech Communication*, S. 297–301, Oct. 2016.

[CHU12] A. Chinaev und R. Haeb-Umbach: „Quality Analysis and Optimization of the MAP-Based Noise Power Spectral Density Tracker“, *In Proc. of the 10th ITG Symposium on Speech Communication*, S. 1–4, Sept. 2012.

[CHU15] A. Chinaev und R. Haeb-Umbach: „On Optimal Smoothing in Minimum Statistics Based Noise Tracking“, *In Proc. of the 16th Annual Interspeech Conf. of the Int. Speech Communication Association (ISCA)*, S. 1785–1789, Sept. 2015.

[CHU16] A. Chinaev und R. Haeb-Umbach: „A Priori SNR Estimation Using a Generalized Decision Directed Approach“, *In Proc. of the 17th Annual Int. Conf. of the Int. Speech Communication Association (ISCA)*, S. 3758–3762, Sept. 2016.

[CHU17] A. Chinaev und R. Haeb-Umbach: „A Generalized Log-Spectral Smplitude Estimator for Single-Channel Speech Enhancement“, *In Proc. of the 42nd IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, S. 4980–4984, Mar. 2017.

[CHUTM13] A. Chinaev, R. Haeb-Umbach, J. Taghia und R. Martin: „Improved Single-Channel Nonstationary Noise Tracking by an Optimized MAP-Based Postprocessor“, *In Proc. of the 38th Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, S. 7477–7481, May 2013.

[CK11] M.-S. Choi und H.-G. Kang: „A Two-Channel Noise Estimator for Speech Enhancement in a Highly Nonstationary Environment“, *IEEE Transactions on Audio, Speech, and Language Processing*, Band 19(4), S. 905–915, May 2011.

[CKTVHU12] A. Chinaev, A. Krueger, D. H. Tran-Vu und R. Haeb-Umbach: „Improved Noise Power Spectral Density Tracking by a MAP-Based Postprocessor“, *In Proc. of the 37th Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, S. 4041–4044, Mar. 2012.

[CN78] R. Curtis und R. J. Niederjohn: „An investigation of several frequency-domain processing methods for enhancing the intelligibility of speech in wideband random noise“, *In Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing.*, Band 3, S. 602–605, Apr. 1978.

[Coh01] I. Cohen: „On speech enhancement under signal presence uncertainty“, *In Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Band 1, S. 661–664, May 2001.

[Coh03] I. Cohen: „Noise Spectrum Estimation in Adverse Environments: Improved Minima Controlled Recursive Averaging“, *IEEE Transactions on Speech and Audio Processing*, Band 11(5), S. 466–475, Sept. 2003.

[Coh04] I. Cohen: „Speech enhancement using a noncausal a priori SNR estimator“, *IEEE Signal Processing Letters*, Band 11(9), S. 725–728, Sept. 2004.

[Coh05] I. Cohen: „Relaxed Statistical Model for Speech Enhancement and a Priori SNR Estimation“, *IEEE Transactions on Speech and Audio Processing*, Band 13(5), S. 870–881, Sept. 2005.

[CPHU14] A. Chinaev, M. Puels und R. Haeb-Umbach: „Spectral Noise Tracking for Improved Nonstationary Noise Robust ASR“, *In Proc. of the 11th ITG Symposium on Speech Communication*, S. 1–4, Sept. 2014.

[CR83] R. E. Crochiere und L. R. Rabiner: *Multirate digital signal processing*, Prentice Hall, 1983.

[Cro80] R. Crochiere: „A weighted overlap-add method of short-time Fourier analysis/synthesis“, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Band 28(1), S. 99–102, Feb. 1980.

[CUH15] D. Clevert, T. Unterthiner und S. Hochreiter: „Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs)“, *CoRR*, Feb. 2015.

[DA09] B. Dashtbozorg und H. R. Abutalebi: „Adaptive MMSE speech spectral amplitude estimator under signal presence uncertainty“, *In Proc. of IEEE European Signal Processing Conference*, S. 209–212, Aug. 2009.

[Dav02] G. M. Davis: *Noise reduction in speech applications*, Band 7, CRC Press, 2002.

[DO03] L. Deng und D. O'Shaughnessy: *Speech processing: a dynamic and optimization-oriented approach*, CRC Press, 2003.

[Dob95] G. Doblinger: „Computationally Efficient Speech Enhancement By Spectral Minima Tracking In Subbands“, *In Proc. of European Conference on Speech Communication and Technology (EUROSPEECH)*, S. 1513–1516, Sept. 1995.

[DP17] B. Das und A. Panda: „Robust Front-End processing for speech recognition in noisy conditions“, *In Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, S. 5235–5239, Mar. 2017.

[DS09] H. Ding und Y. Soon: „An adaptive time-shift analysis for DCT based speech enhancement“, *In Proc. of the 7th Int. Conf. on Information, Communications and Signal Processing*, S. 1–4, Dec. 2009.

[DTI05] T. H. Dat, K. Takeda und F. Itakura: „Generalized gamma modeling of speech and its online estimation for speech enhancement“, *In Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Band 4, S. IV–181, Mar. 2005.

[EH08] J. Erkelens und R. Heusdens: „Fast noise tracking based on recursive smoothing of MMSE noise power estimates“, *In Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, S. 4873–4876, Mar. 2008.

[EHHJ07] J. S. Erkelens, R. C. Hendriks, R. Heusdens und J. Jensen: „Minimum mean-square error estimation of discrete Fourier coefficients with generalized Gamma priors“, *IEEE Transactions on Audio, Speech, and Language Processing*, Band 15(6), S. 1741–1752, Aug. 2007.

[EJH07] J. Erkelens, J. Jensen und R. Heusdens: „A data-driven approach to optimizing spectral speech enhancement methods for various error criteria“, *Speech Communication*, Band 49(7), S. 530–541, Aug. 2007.

[EM83] Y. Ephraim und D. Malah: „Speech enhancement using optimal non-linear spectral amplitude estimation“, *In Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Band 8, S. 1118–1121, Apr. 1983.

[EM84] Y. Ephraim und D. Malah: „Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator“, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Band ASSP-32(6), S. 1109–1121, Dec. 1984.

[EM85] Y. Ephraim und D. Malah: „Speech Enhancement Using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator“, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Band 33(2), S. 443–445, Apr. 1985.

[EMTF15] S. Elshamy, N. Madhu, W. Tirry und T. Fingscheidt: „An Iterative Speech Model-Based A Priori SNR Estimator“, *In Proc. of the 16th Annual Conf. of the Int. Speech Communication Association*, S. 1740–1744, Sept. 2015.

[FBR04] N. Fan, R. V. Balan und J. Rosca: „Comparison of wavelet-and FFT-based single-channel speech signal noise reduction techniques“, *Optics East*, S. 127–138, Int. Society for Optics and Photonics, 2004.

[FDGM86] W. M. Fisher, G. R. Doddington und K. M. Goudie-Marshall: „The DARPA speech recognition research database: specifications and status“, *In Proc. of DARPA Workshop on Speech Recognition*, S. 93–99, Feb. 1986.

[FF12] B. Fodor und T. Fingscheidt: „MMSE speech enhancement under speech presence uncertainty assuming (generalized) gamma speech priors throughout“, *In Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, S. 4033–4036, Mar. 2012.

[FG50] H. Fletcher und R. H. Galt: „The perception of speech and its relation to telephony“, *The Journal of the Acoustical Society of America*, Band 22(2), S. 89–151, Mar. 1950.

[FG14] B. Fodor und T. Gerkmann: „A posteriori speech presence probability estimation based on averaged observations and a super-Gaussian speech model“, *In Proc. of Int. Workshop on Acoustic Signal Enhancement (IWAENC)*, S. 11–15, Sept. 2014.

[Fis96] G. S. Fishman: *Monte Carlo: Concepts, Algorithms and Applications*, Springer, 1996.

[FRB07] N. Fan, J. Rosca und R. Balan: „Speech Noise Estimation using Enhanced Minima Controlled Recursive Averaging“, *In Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, S. IV–581–IV–584, June 2007.

[FSS08] T. Fingscheidt, S. Suhadi und K. Steinert: „Towards objective quality assessment of speech enhancement systems in a black box approach“, *In Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, S. 273–276, Feb. 2008.

[FW10] Z. H. Fu und J. F. Wang: „Speech presence probability estimation based on integrated time-frequency minimum tracking for speech enhancement in adverse environments“, *In Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, S. 4258–4261, Mar. 2010.

[GB10] X. Glorot und Y. Bengio: „Understanding the difficulty of training deep feedforward neural networks“, *In Proc. of Int. Conf. on Artificial Intelligence and Statistics (AISTATS)*, May 2010.

[GBGM80] R. Gray, A. Buzo, A. Gray und Y. Matsuyama: „Distortion measures for speech processing“, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Band 28(4), S. 367–376, Aug. 1980.

[GBM08] T. Gerkmann, C. Breithaupt und R. Martin: „Improved A Posteriori Speech Presence Probability Estimation Based on a Likelihood Ratio With Fixed Priors“, *IEEE Transactions on Audio, Speech, and Language Processing*, Band 16(5), S. 910–919, July 2008.

[GGPP07] J. Garofolo, D. Graff, D. Paul und D. Pallett: „CSI-I (WSJO) complete“, *Linguistic Data Consortium, Philadelphia*, 2007.

[GH11] T. Gerkmann und R. C. Hendriks: „Noise power estimation based on the probability of speech presence“, *In Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, S. 145–148, Oct. 2011.

[GH12] T. Gerkmann und R. C. Hendriks: „Unbiased MMSE-Based Noise Power Estimation With Low Complexity and Low Tracking Delay“, *IEEE Transactions on Audio, Speech, and Language Processing*, Band 20(4), S. 1383–1393, May 2012.

[GKR12] T. Gerkmann, M. Krawczyk und R. Rehr: „Phase estimation in speech enhancement; Unimportant, important, or impossible?“, *In Proc. of IEEE Convention of Electrical Electronics Engineers in Israel (IEEEI)*, S. 1–5, Nov. 2012.

[Gla15] T. Glarner: *Lern- und Inferenzverfahren für symmetrische 2D-HMMs zur Sprachpräsenzdetektion*, Master's thesis, Sept. 2015.

[GLF+93] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett und N. L. Dahlgren: „DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus CDROM", Feb. 1993.

[GM76] A. Gray und J. Markel: „Distance measures for speech processing", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Band 24(5), S. 380–391, Oct. 1976.

[GMV96] S. Gustafsson, R. Martin und P. Vary: „On the optimization of speech enhancement systems using instrumental measures", *In Proc. of Workshop on Quality Assessment in Speech, Audio, and Image Communication*, S. 36–40, Mar. 1996.

[GTT98] Z. Goh, K.-C. Tan und T. G. Tan: „Postprocessing method for suppressing musical noise generated by spectral subtraction", *IEEE Transactions on Speech and Audio Processing*, Band 6(3), S. 287–292, May 1998.

[HCS+07] A. Hussain, M. Chetouani, S. Squartini, A. Bastari und F. Piazza: *Progress in Nonlinear Speech Processing*, chap. Nonlinear Speech Enhancement: An Overview, S. 217–248, Springer Berlin Heidelberg, Berlin, Heidelberg, 2007.

[HDHU16] J. Heymann, L. Drude und R. Haeb-Umbach: „Neural network based spectral mask estimation for acoustic beamforming", *In Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, S. 196–200, Mar. 2016.

[HDTC12] A. Hirszhorn, D. Dov, R. Talmon und I. Cohen: „Transient Interference Suppression in Speech Signals Based on the OM-LSA Algorithm", *In Proc. of Int. Workshop on Acoustic Signal Enhancement*, S. 1–4, Sept. 2012.

[HE95] H. Hirsch und C. Ehrlicher: „Noise estimation techniques for robust speech recognition", *In Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Band 1, S. 153–156, May 1995.

[HHJ10] R. C. Hendriks, R. Heusdens und J. Jensen: „MMSE based noise PSD tracking with low complexity", *In Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, S. 4266–4269, Mar. 2010.

[HHJK09] R. C. Hendriks, R. Heusdens, J. Jensen und U. Kjems: „Fast noise PSD estimation with low complexity", *In Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, S. 3881–3884, Apr. 2009.

[Hir93] H. G. Hirsch: „Estimation of noise spectrum and its application to SNR-estimation and speech enhancement", Int. Computer Science Institute, Dec. 1993.

[HJH08] R. C. Hendriks, J. Jensen und R. Heusdens: „Noise tracking using DFT domain subspace decompositions“, *IEEE Transactions on Audio, Speech and Language Processing*, Band 16(3), S. 541–553, Mar. 2008.

[HL06] Y. Hu und P. Loizou: „Subjective Comparison of Speech Enhancement Algorithms“, *In Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Band 1, S. I–I, May 2006.

[HL07] Y. Hu und P. C. Loizou: „A comparative intelligibility study of single-microphone noise reduction algorithms“, *The Journal of the Acoustical Society of America*, Band 122(3), S. 1777–1786, Sept. 2007.

[HL08] Y. Hu und P. C. Loizou: „Evaluation of objective quality measures for speech enhancement“, *IEEE Transactions on Audio, Speech, and Language Processing*, Band 16(1), S. 229–238, Jan. 2008.

[HOT06] G. E. Hinton, S. Osindero und Y.-W. Teh: „A fast learning algorithm for deep belief nets“, *Neural computation*, Band 18(7), S. 1527–1554, July 2006.

[HS97] S. Hochreiter und J. Schmidhuber: „Long short-term memory“, *Neural computation*, Band 9(8), S. 1735–1780, Nov. 1997.

[HS06] G. E. Hinton und R. R. Salakhutdinov: „Reducing the dimensionality of data with neural networks“, *Science*, Band 313(5786), S. 504–507, July 2006.

[HSK04] M. K. Hasan, S. Salahuddin und M. R. Khan: „A modified a priori SNR for speech enhancement using spectral subtraction rules“, *IEEE Signal Processing Letters*, Band 8(4), S. 450–453, Apr. 2004.

[HSW89] K. Hornik, M. Stinchcombe und H. White: „Multilayer feedforward networks are universal approximators“, *Neural networks*, Band 2(5), S. 359–366, July 1989.

[IS06] B. Iser und G. Schmidt: „Bewertung verschiedener Methoden zur Erzeugung des Anregungssignals innerhalb eines Algorithmus zur Bandbreitenerweiterung“, *In Proc. of ITG Fachtagung Sprachkommunikation*, Kiel (Germany), Apr. 2006.

[IS08] B. Iser und G. Schmidt: „Receive Side Processing for Automotive Hands-Free Systems“, *Hands-Free Speech Communication and Microphone Arrays*, S. 236–239, May 2008.

[IS15] S. Ioffe und C. Szegedy: „Batch normalization: Accelerating deep network training by reducing internal covariate shift“, *arXiv e-prints*, Mar. 2015.

[IST+11] T. Inoue, H. Saruwatari, Y. Takahashi, K. Shikano und K. Kondo: „Theoretical Analysis of Musical Noise in Generalized Spectral Subtraction Based on Higher Order Statistics“, *In Proc. of IEEE Transactions on Audio, Speech, and Language Processing*, Band 19(6), S. 1770–1779, Aug. 2011.

[ITS+10] T. Inoue, Y. Takahashi, H. Saruwatari, K. Shikano und K. Rondo: „Theoretical analysis of musical noise in generalized spectral subtraction: Why should not use power/amplitude subtraction?“, *In Proc. of European Signal Processing Conference*, S. 994–998, Aug. 2010.

[ITU96] ITU: „P.800: Methods for subjective determination of transmission quality“, *Int. Telecommunication Union Recommendation*, (P.800), Aug. 1996.

[ITU01] ITU: „P.862: Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs“, *Int. Telecommunication Union Recommendation*, Feb. 2001.

[ITU03] ITU: „P.862.1: Mapping function for transforming P. 862 raw result scores to MOS-LQO“, *Int. Telecommunication Union Recommendation*, Nov. 2003.

[ITU07] ITU: „P.862.3: Application guide for objective quality measurement based on recommendations P.862, P.862.1 and P.862.2“, *Int. Telecommunication Union Recommendation*, Nov. 2007.

[JBHH05] J. Jensen, I. Batina, R. C. Hendriks und R. Heusdens: „A study of the distribution of time-domain speech samples and discrete Fourier coefficients“, *In Proc. of DSP Valley's Annual Research and Technology Symposium (DARTS)*, Band 1, S. 155–158, 2005.

[JH12] J. Jensen und R. C. Hendriks: „Spectral magnitude minimum mean-square error estimation using binary and continuous gain functions“, *IEEE Transactions on Audio, Speech, and Language Processing*, Band 20(1), S. 92–102, Jan. 2012.

[JS93] D. Johnson und P. N. Shami: „The signal processing information base“, *IEEE Signal Processing Magazine*, Band 10, S. 36–43, Oct. 1993.

[KA10] M. Khodabin und A. Ahmadabadi: „Some properties of generalized gamma distribution“, *Mathematical Sciences*, Band 4(1), S. 9–28, Mar. 2010.

[Kay93] S. M. Kay: *Fundamentals of statistical signal processing*, Prentice Hall PTR, 1993.

[KC00] N. S. Kim und J.-H. Chang: „Spectral enhancement based on global soft decision“, *IEEE Signal Processing Letters*, Band 7(5), S. 108–110, May 2000.

[KGW+89] W. M. Kushner, V. Goncharoff, C. Wu, V. Nguyen und J. N. Damoulakis: „The effects of subtractive-type speech enhancement/noise reduction algorithms on parameter estimation for improved recognition and coding in high noise environments“, *In Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, S. 211–214, May 1989.

[KHU11] A. Krueger und R. Haeb-Umbach: „MAP-based estimation of the parameters of non-stationary Gaussian processes from noisy observations“, *In Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, S. 3596–3599, May 2011.

[KLHL09] G. Kim, Y. Lu, Y. Hu und P. C. Loizou: „An algorithm that improves speech intelligibility in noise for normal-hearing listeners“, *The Journal of the Acoustical Society of America*, Band 126(3), S. 1486–1494, Sept. 2009.

[KPC09] J. M. Kum, Y. S. Park und J. H. Chang: „Speech enhancement based on minima controlled recursive averaging incorporating conditional maximum a posteriori criterion“, *In Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, S. 4417–4420, Apr. 2009.

[KSS06] M. Kato, A. Sugiyama und M. Serizawa: „Noise suppression with high speech quality based on weighted noise estimation and MMSE STSA“, *Electronics and Communications in Japan (Part III: Fundamental Electronic Science)*, Band 89(2), S. 43–53, July 2006.

[LBH15] Y. LeCun, Y. Bengio und G. Hinton: „Deep learning“, *Nature*, Band 521(7553), S. 436–444, May 2015.

[LFJA09] J. Li, Q.-J. Fu, H. Jiang und M. Akagi: „Psychoacoustically-motivated adaptive β-order generalized spectral subtraction for cochlear implant patients“, *In Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, S. 4665–4668, Apr. 2009.

[Lim78] J. Lim: „Evaluation of a correlation subtraction method for enhancing speech degraded by additive white noise“, *IEEE Transactions on Acoustics, Speech and Signal Processing*, Band 26(5), S. 471–472, Oct. 1978.

[LK11] P. C. Loizou und G. Kim: „Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions“, *IEEE Transactions on Audio, Speech, and Language Processing*, Band 19(1), S. 47–56, Jan. 2011.

[LKS89] L. F. Lamel, R. H. Kassel und S. Seneff: „Speech database development: Design and analysis of the acoustic-phonetic corpus“, *Speech Input/Output Assessment and Speech Databases*, Sept. 1989.

[LL08] Y. Lu und P. C. Loizou: „A geometric approach to spectral subtraction“, *Speech Communication*, Band 50(6), S. 453–466, June 2008.

[LL11] Y. Lu und P. C. Loizou: „Estimators of the Magnitude-Squared Spectrum and Methods for Incorporating SNR Uncertainty“, *IEEE Transactions on Audio, Speech, and Language Processing*, Band 19(5), S. 1123–1137, July 2011.

[LLC14] S. Lee, C. Lim und J.-H. Chang: „A new a priori SNR estimator based on multiple linear regression technique for speech enhancement“, *Digital Signal Processing*, Band 30, S. 154–164, Apr. 2014.

[LO79] J. Lim und A. Oppenheim: „Enhancement and bandwidth compression of noisy speech“, *Proceedings of the IEEE*, Band 67(12), S. 1586–1604, Dec. 1979.

[Loi13] P. C. Loizou: *Speech enhancement: Theory and Practice*, CRC Press, Jan. 2013.

[LSH+08] J. Li, S. Sakamoto, S. Hongo, M. Akagi und Y. Suzuki: „Adaptive β-order generalized spectral subtraction for speech enhancement“, *Signal Processing*, Band 88(11), S. 2764–2776, June 2008.

[LSTW14] C. Lüke, G. Schmidt, A. Theiß und J. Withopf: „In-Car Communication“, *Smart Mobile In-Vehicle Systems*, S. 97–118, Springer, Sept. 2014.

[LTMH13] X. Lu, Y. Tsao, S. Matsuda und C. Hori: „Speech enhancement based on deep denoising autoencoder.“, *In Proc. of INTERSPEECH Conference*, S. 436–440, Aug. 2013.

[LV05] T. Lotter und P. Vary: „Speech enhancement by MAP spectral amplitude estimation using a super-Gaussian speech model“, *EURASIP Journal on Applied Signal Processing*, S. 1110–1126, 2005.

[LYZ+11] J. Li, L. Yang, J. Zhang, Y. Yan, Y. Hu, M. Akagi und P. C. Loizou: „Comparative intelligibility investigation of single-channel noise-reduction algorithms for Chinese, Japanese, and Englisha)“, *The Journal of the Acoustical Society of America*, Band 129(5), S. 3291–3301, May 2011.

[MA+16] H. Momeni, H. R. Abutalebi *et al.*: „Conditional MMSE-based single-channel speech enhancement using inter-frame and inter-band correlations“, *In Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, S. 5215–5219, Mar. 2016.

[Mar90] C. Marvin: *When old technologies were new: Thinking about electric communication in the late nineteenth century*, {Oxford University Press, USA}, 1990.

[Mar94] R. Martin: „Spectral Subtraction Based on Minimum Statistics“, *In Proc. of EUSIPCO Conference*, S. 1182–1185, Sept. 1994.

[Mar01] R. Martin: „Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics“, *IEEE Transactions on Speech and Audio Processing*, Band 9(5), S. 504–512, July 2001.

[Mar03] R. Martin: „Statistical methods for the enhancement of noisy speech“, *In Proc. of IEEE Int. Workshop on Acoustic Echo and Noise Control*, S. 43–65, Springer Verlag, 2003.

[Mar05] R. Martin: „Speech enhancement based on minimum mean-square error estimation and supergaussian priors“, *IEEE Transactions on Speech and Audio Processing*, Band 13(5), S. 845–856, Sept. 2005.

[MCA99] D. Malah, R. V. Cox und A. J. Accardi: „Tracking speech-presence uncertainty to improve speech enhancement in non-stationary noise environments“, *In Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Band 2, S. 789–792, Mar. 1999.

[ME68] D. Middleton und R. Esposito: „Simultaneous optimum detection and estimation of signals in noise“, *IEEE Transactions on Information Theory*, Band 14(3), S. 434–444, May 1968.

[MHA14] H. Momeni, E. Habets und H. Abutalebi: „Single-channel speech presence probability estimation using inter-frame and inter-band correlations“, *In Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, S. 2903–2907, May 2014.

[MHL09] J. Ma, Y. Hu und P. C. Loizou: „Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions“, *The Journal of the Acoustical Society of America*, Band 125(5), S. 3387–3405, May 2009.

[ML01] R. Martin und T. Lotter: „Optimal recursive smoothing of non-stationary periodograms“, *In Proc. of Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, S. 167–170, Sept. 2001.

[MM80] R. McAulay und M. Malpass: „Speech enhancement using a soft-decision noise suppression filter“, *IEEE Transactions on Acoustics, Speech and Signal Processing*, Band 28(2), S. 137–145, Apr. 1980.

[MMCA04] R. Martin, D. Malah, R. V. Cox und A. J. Accardi: „A noise reduction preprocessor for mobile voice communication“, *EURASIP Journal on Applied Signal Processing*, Band 2004, S. 1046–1058, July 2004.

[MSS16] P. Mowlaee, R. Saeidi und Y. Stylianou: „Advances in phase-aware signal processing in speech communication“, *Speech Communication*, Band 81, S. 1–29, May 2016.

[MT16] S. Mirsamadi und I. Tashev: „Causal Speech Enhancement Combining Data-Driven Learning and Suppression Rule Estimation“, *In Proc. of INTERSPEECH Conference*, S. 2870–2874, Sept. 2016.

[MWJ00] R. Martin, I. Wittke und P. Jax: „Optimized estimation of spectral parameters for the coding of noisy speech“, *In Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Band 3, S. 1479–1482, June 2000.

[Nim92] B. Nimens: „Sound ideas: sound effects collection. Ser. 6000“, 1992, [URL] https://www.sound-ideas.com.

[NW13] A. Narayanan und D. Wang: „Ideal ratio mask estimation using deep neural networks for robust speech recognition“, *In Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, S. 7092–7096, May 2013.

[PC07a] Y.-S. Park und J.-H. Chang: „A novel approach to a robust a priori SNR estimator in speech enhancement“, *IEICE Transactions on Communications*, Band 90(8), S. 2182–2185, Aug. 2007.

[PC07b] E. Plourde und B. Champagne: „Further analysis of the β-order MMSE STSA estimator for speech enhancement“, *In Proc. of Canadian Conference on Electrical and Computer Engineering*, S. 1594–1597, 2007.

[PG04] S. Parveen und P. Green: „Speech enhancement with missing data techniques using recurrent neural networks“, *In Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Band 1, S. I–733, May 2004.

[Pin87] F. J. Pineda: „Generalization of back-propagation to recurrent neural networks“, *Physical review letters*, Band 59(19), S. 2229, Nov. 1987.

[PJHUF16] A. Plinge, F. Jacob, R. Haeb-Umbach und G. A. Fink: „Acoustic Microphone Geometry Calibration - An overview and experimental evaluation of state-of-the-art algorithms“, *IEEE Signal Processing Magazine*, July 2016.

[PMS06] C. Plapous, C. Marro und P. Scalart: „Improved Signal-to-Noise Ratio Estimation for Speech Enhancement“, *IEEE Transactions on Audio, Speech, and Language Processing*, Band 14(6), S. 2098–2108, Nov. 2006.

[PP02] A. Papoulis und S. U. Pillai: *Probability, random variables, and stochastic processes*, Tata McGraw-Hill Education, 2002.

[Pud99] H. Puder: „Single channel noise reduction using time-frequency dependent voice activity detection“, *In Proc. of Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, S. 68–71, Sept. 1999.

[PWS10] K. Paliwal, K. Wojcicki und B. Schwerin: „Single-channel speech enhancement using spectral subtraction in the short-time modulation domain“, *Speech Communication*, Band 52(5), S. 450 – 475, May 2010.

[PWS11] K. Paliwal, K. Wójcicki und B. Shannon: „The importance of phase in speech enhancement“, *Speech Communication*, Band 53(4), S. 465–494, Dec. 2011.

[QBC88] S. R. Quackenbush, T. P. Barnwell und M. A. Clements: *Objective measures of speech quality*, Prentice Hall, 1988.

[RAS06] S. Ravindran, D. V. Anderson und M. Slaney: „Improving the noise-robustness of mel-frequency cepstral coefficients for speech processing“, *Reconstruction*, Band 12, S. 14, Sept. 2006.

[RBKS14] V. K. Rajan, C. Baasch, M. Krini und G. Schmidt: „Improvement in Listener Comfort Through Noise Shaping Using a Modified Wiener Filter Approach“, *In Proc. of the 11. ITG Symposium on Speech Communication*, S. 1–4, Sept. 2014.

[RBQ+08] D. Rudoy, P. Basu, T. F. Quatieri, B. Dunn und P. J. Wolfe: „Adaptive short-time analysis-synthesis for speech enhancement“, *In Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, S. 4905–4908, Mar. 2008.

[RHW86] D. E. Rumelhart, G. E. Hinton und R. J. Williams: „Learning representations by back-propagating errors“, *Nature*, Band 323, S. 533–536, Oct. 1986.

[Ric48] S. Rice: „Statistical properties of a sine wave plus random noise“, *Bell System Technical Journal, The*, Band 27(1), S. 109–157, Jan. 1948.

[RKRS16] V. K. Rajan, M. Krini, K. Rodemer und G. Schmidt: „Signal processing techniques for seat belt microphone arrays“, *EURASIP Journal on Advances in Signal Processing*, (1), S. 35, Mar. 2016.

[RL06] S. Rangachari und P. C. Loizou: „A noise-estimation algorithm for highly non-stationary environments“, *Speech Communication*, Band 48(2), S. 220–231, Feb. 2006.

[RMN09] R. Raina, A. Madhavan und A. Y. Ng: „Large-scale deep unsupervised learning using graphics processors“, *In Proc. of the 26th annual Int. Conf. on Machine Learning*, S. 873–880, ACM, June 2009.

[Rou73] D. N. Rousu: „Weibull skewness and kurtosis as a function of the shape parameter“, *Technometrics*, S. 927–930, Nov. 1973.

[RS07] L. R. Rabiner und R. W. Schafer: *Introduction to digital speech processing*, Band 1, Now Publishers Inc., 2007.

[RSU00] T. Richardson, A. Shokrollahi und R. Urbanke: „Design of provably good low-density parity check codes“, *In Proc. of IEEE Int. Symposium on Information Theory*, S. 199, June 2000.

[SA05] K. V. Sørensen und S. V. Andersen: „Speech enhancement with natural sounding residual noise based on connected time-frequency speech presence regions“, *EURASIP Journal on Advances in Signal Processing*, (18), S. 1–11, Mar. 2005.

[Sch65] M. R. Schroeder: „Apparatus for suppressing noise and distortion in communication signals“, Apr. 1965.

[Sch10] J. Schmalenströer: *Akustische Szenenanalyse für die ambiente Kommunikation im vernetzten Haus.*, Ph.D. thesis, University of Paderborn, Mar. 2010.

[SF96] P. Scalart und J. Filho: „Speech enhancement based on a priori signal to noise estimation“, *In Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Band 2, S. 629–632 vol. 2, May 1996.

[SG88] H. J. Steeneken und F. W. Geurtsen: „Description of the RSG-10 noise database“, *report IZF*, Band 3, 1988.

[SHK+14] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever und R. Salakhutdinov: „Dropout: A Simple Way to Prevent Neural Networks from Overfitting“, Band 15, S. 1929–1958, June 2014.

[SK00] I. Y. Soon und S. N. Koh: „Low distortion speech enhancement“, *IEE Proc. of Vision, Image and Signal Processing*, Band 147(3), S. 247–253, June 2000.

[SKS99] J. Sohn, N. S. Kim und W. Sung: „A statistical model-based voice activity detection“, *IEEE Signal Processing Letters*, Band 6(1), S. 1–3, Jan. 1999.

[SKY98] Y. Soon, S. N. Koh und C. K. Yeo: „Noisy speech enhancement using discrete cosine transform“, *Speech Communication*, Band 24(3), S. 249–257, June 1998.

[SKY99] Y. Soon, S. N. Koh und C. K. Yeo: „Improved noise suppression filter using self-adaptive estimator of probability of speech absence“, *Signal Processing*, Band 75(2), S. 151–159, June 1999.

[SLF11] S. Suhadi, C. Last und T. Fingscheidt: „A Data-Driven Approach to A Priori SNR Estimation“, *IEEE Transactions on Audio, Speech, and Language Processing*, Band 19(1), S. 186–195, Jan. 2011.

[SMP13] A. Sugiyama, R. Miyahara und K. Park: „Impact-noise suppression with phase-based detection“, *In Proc. of the European Signal Processing Conference (EUSIPCO).*, S. 1–5, Sept. 2013.

[SOLG14] H. Sun, S. Ou, R. Liu und Y. Gao: „A variable momentum factor algorithm for a priori SNR estimation in speech enhancement“, *In Proc. of the 7th Int. Congress on Image and Signal Processing (CISP)*, S. 888–892, Oct. 2014.

[SSB+15] B. Schuller, S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, F. Burkhardt, R. van Son, F. Weninger, F. Eyben, T. Bocklet *et al.*: „A Survey on perceived speaker traits: Personality, likability, pathology, and the first challenge“, *Computer Speech & Language*, Band 29(1), S. 100–131, Jan. 2015.

[STCT98] B. L. Sim, Y. C. Tong, J. S. Chang und C. T. Tan: „A parametric formulation of the generalized spectral subtraction method“, *IEEE Transactions on Speech and Audio Processing*, Band 6(4), S. 328–337, July 1998.

[STWJ13] Y.-C. Su, Y. Tsao, J.-E. Wu und F.-R. Jean: „Speech enhancement using generalized maximum a posteriori spectral amplitude estimator“, *In Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, S. 7467–7471, May 2013.

[SV06] B. Sauert und P. Vary: „Near end listening enhancement: Speech intelligibility improvement in noisy environments“, *In Proc. of IEEE Int. Conf. on Acoustics Speech and Signal Processing Proceedings*, Band 1, S. 493–496, May 2006.

[SYMA05] R. Saab, O. Yilmaz, M. J. McKeown und R. Abugharbieh: „Underdetermined sparse blind source separation with delays“, *In Proc. of the 1st Workshop on Signal Processing with Sparse/Structured Representations*, S. 67–70, July 2005.

[TA10] I. Tashev und A. Acero: „Statistical modeling of the speech signal“, *In Proc. of Int. Workshop on Acoustic, Echo, and Noise Control*, Aug. 2010.

[TCG13] R. Talmon, I. Cohen und S. Gannot: „Single-channel transient interference suppression with diffusion maps“, *IEEE Transactions on Audio, Speech, and Language Processing*, Band 21(1), S. 132–144, Jan. 2013.

[THHJ10] C. H. Taal, R. C. Hendriks, R. Heusdens und J. Jensen: „A short-time objective intelligibility measure for time-frequency weighted noisy speech.“, *In Proc. of IEEE Int. Conf. on Acoustics Speech and Signal Processing*, S. 4214–4217, Mar. 2010.

[THHJ11] C. Taal, R. Hendriks, R. Heusdens und J. Jensen: „An Algorithm for Intelligibility Prediction of Time-Frequency Weighted Noisy Speech“, *IEEE Transactions on Audio, Speech, and Language Processing*, Band 19(7), S. 2125–2136, Sept. 2011.

[TJB72] Y. Takefuta, E. Jancosek und M. Brunt: „A statistical analysis of melody curves in the intonation of American English“, *In Proc. of the 7th Int. Congress of Phonetic Sciences*, S. 1035–1039, Aug. 1972.

[TL16] Y. Tsao und Y.-H. Lai: „Generalized maximum a posteriori spectral amplitude estimation for speech enhancement“, *Speech Communication*, Band 76, S. 112 – 126, Feb. 2016.

[TLA09] I. Tashev, A. Lovitt und A. Acero: „Unified framework for single channel speech enhancement“, *In Proc. of IEEE Pacific Rim Conference on Communications, Computers and Signal Processing*, S. 883–888, Aug. 2009.

[TTM+11] J. Taghia, J. Taghia, N. Mohammadiha, J. Sang, V. Bouse und R. Martin: „An evaluation of noise power spectral density estimation algorithms in adverse acoustic environments“, *In Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, S. 4640–4643, May 2011.

[TVHU13] D. H. Tran Vu und R. Haeb-Umbach: „Using the turbo principle for exploiting temporal and spectral correlations in speech presence probability estimation“, *In Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP),*, S. 863–867, May 2013.

[Vas08] S. V. Vaseghi: *Advanced digital signal processing and noise reduction*, John Wiley & Sons, 2008.

[VHH98] P. Vary, U. Heute und W. Hess: *Digitale Sprachsignalverarbeitung*, Stuttgart: Teubner,, 1998.

[VM06] P. Vary und R. Martin: *Digital speech transmission: Enhancement, coding and error concealment*, John Wiley & Sons, 2006.

[Vor15] S. Voran: „Exploration of the additivity approximation for spectral magnitudes“, *In Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, S. 1–5, Oct. 2015.

[VS93] A. Varga und H. J. M. Steeneken: „Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems“, *Speech Communication*, Band 12(3), S. 247–251, July 1993.

[Wan05] D. Wang: „On ideal binary mask as the computational goal of auditory scene analysis“, *Speech separation by humans and machines*, S. 181–197, Springer, 2005.

[WAP75] M. R. Weiss, E. Aschkenasy und T. W. Parsons: „Study and development of the INTEL technique for improving speech intelligibility“, Tech. Rep., Apr. 1975.

[Wei51] W. Weibull: „A Statistical Distribution Function of Wide Applicability“, *Journal of Applied Mechanics*, Band 18, S. 293–297, Sept. 1951.

[Wei09] T. Weickert: *Nichtstationäre Filterung mit Hilfe analytischer wavelet packets am Beispiel von Sprachsignalen*, Univ.-Verlag Karlsruhe, 2009.

[WHU07] E. Warsitz und R. Haeb-Umbach: „Blind acoustic beamforming based on generalized eigenvalue decomposition“, *IEEE Transactions on Audio, Speech, and Language Processing*, Band 15(5), S. 1529–1539, July 2007.

[WL82] D. Wang und J. Lim: „The unimportance of phase in speech enhancement“, *IEEE Transactions on Acoustics, Speech and Signal Processing*, Band 30(4), S. 679–681, Aug. 1982.

[WN99] E. A. Wan und A. T. Nelson: „Networks for speech enhancement“, *Handbook of Neural Networks for Speech Processing. Artech House, Boston, USA*, Aug. 1999.

[Wol17] N. R. Wolf: „Schätzung der Sprachpräsenzwahrscheinlichkeit im Bereich generalisierter a posteriori SNR“, Bachelor's thesis, May 2017.

[WR90] J. Wexler und S. Raz: „Discrete Gabor expansions“, *Signal processing*, Band 21(3), S. 207–220, Nov. 1990.

[WSST12] R. Wakisaka, H. Saruwatari, K. Shikano und T. Takatani: „Speech prior estimation for generalized minimum mean-square error short-time spectral amplitude estimator“, *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, Band 95(2), S. 591–595, Feb. 2012.

[XDDL14] Y. Xu, J. Du, L. R. Dai und C. H. Lee: „An Experimental Study on Speech Enhancement Based on Deep Neural Networks“, *IEEE Signal Processing Letters*, Band 21(1), S. 65–68, Jan. 2014.

[XVC93] F. Xie und D. Van Compernolle: „Speech enhancement by nonlinear spectral estimation-a unifying approach.“, *In Proc. of EUROSPEECH Conference*, Band 93, S. 617–620, Sept. 1993.

[Yan93] J. Yang: „Frequency domain noise suppression approaches in mobile telephone systems“, *In Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Band 2, S. 363–366, Apr. 1993.

[YF11] H. Yu und T. Fingscheidt: „A figure of merit for instrumental optimization of noise reduction algorithms“, *In Proc. of the 5th Biennial Workshop on Digital Signal Processing for In-Vehicle Systems*, Sept. 2011.

[YKR03] C. You, S. Koh und S. Rahardja: „Adaptive β-order MMSE estimation for speech enhancement“, *In Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Band 1, S. I–900, Apr. 2003.

[YKR05] C. H. You, S. N. Koh und S. Rahardja: „β-order MMSE spectral amplitude estimation for speech enhancement“, *IEEE Transactions on Speech and Audio Processing*, Band 13(4), S. 475–486, July 2005.

[YND13] P. C. Yong, S. Nordholm und H. H. Dam: „Optimization and evaluation of sigmoid function with a priori {SNR} estimate for real-time speech enhancement“, *Speech Communication*, Band 55(2), S. 358 – 376, Sept. 2013.

[YS05] K. Yamashita und T. Shimamura: „Nonstationary noise estimation using low-frequency regions for spectral subtraction“, *IEEE Signal Processing Letters*, Band 12(6), S. 465–468, June 2005.

[Yu09] R. Yu: „A low-complexity noise estimation algorithm based on smoothing of noise power estimation and estimation bias correction“, *In Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, S. 4421–4424, Apr. 2009.

[Yu13] H. Yu: *Post-Filter Optimization for Multichannel Automotive Speech Enhancement*, Ph.D. thesis, Juli 2013.

[YZZ16] R. Yao, Z. Zeng und P. Zhu: „A priori SNR estimation and noise estimation for speech enhancement“, *EURASIP Journal on Advances in Signal Processing*, (1), S. 101, Sept. 2016.

[ZKS12] T.-C. Zorilua, V. Kandia und Y. Stylianou: „Speech-in-noise intelligibility improvement based on power recovery and dynamic range compression“, *In Proc. of the 20th European Signal Processing Conference (EUSIPCO)*, S. 2075–2079, Aug. 2012.

[ZSG90] V. Zue, S. Seneff und J. Glass: „Speech database development at MIT: TIMIT and beyond“, *Speech Communication*, Band 9(4), S. 351–356, Aug. 1990.

[ZSV14] W. Zaremba, I. Sutskever und O. Vinyals: „Recurrent Neural Network Regularization“, *Computing Research Repository (CoRR)*, Sept. 2014.

[ZW13] X. L. Zhang und J. Wu: „Denoising deep neural networks based voice activity detection“, *In Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, S. 853–857, May 2013.

[ZZZLy12] G. Zhao, B. Zhou, X. Zhang und S. Lu-ying: „A new speech enhancement algorithm with generalized Gamma speech model“, *In Proc. of IEEE Int. Conf. on Wireless Communications & Signal Processing (WCSP)*, S. 1–5, Oct. 2012.

Publikationen mit eigener Beteiligung

[CHDHU16] A. Chinaev, J. Heymann, L. Drude und R. Haeb-Umbach: „Noise-Presence-Probability-Based Noise PSD Estimation by Using DNNs“, *In Proc. of the 12th ITG Symposium on Speech Communication*, S. 26–30, Oct. 2016.

[CHHU16] A. Chinaev, J. Heitkaemper und R. Haeb-Umbach: „A Priori SNR Estimation Using Weibull Mixture Model“, *In Proc. of the 12th ITG Symposium on Speech Communication*, S. 297–301, Oct. 2016.

[CHU12] A. Chinaev und R. Haeb-Umbach: „Quality Analysis and Optimization of the MAP-Based Noise Power Spectral Density Tracker“, *In Proc. of the 10th ITG Symposium on Speech Communication*, S. 1–4, Sept. 2012.

[CHU13] A. Chinaev und R. Haeb-Umbach: „MAP-based Estimation of the Parameters of a Gaussian Mixture Model in the Presence of Noisy Observations“, *In Proc. of the 38th Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, S. 3352–3356, May 2013.

[CHU15] A. Chinaev und R. Haeb-Umbach: „On Optimal Smoothing in Minimum Statistics Based Noise Tracking“, *In Proc. of the 16th Annual Interspeech Conf. of the Int. Speech Communication Association (ISCA)*, S. 1785–1789, Sept. 2015.

[CHU16] A. Chinaev und R. Haeb-Umbach: „A Priori SNR Estimation Using a Generalized Decision Directed Approach“, *In Proc. of the 17th Annual Int. Conf. of the Int. Speech Communication Association (ISCA)*, S. 3758–3762, Sept. 2016.

[CHU17] A. Chinaev und R. Haeb-Umbach: „A Generalized Log-Spectral Smplitude Estimator for Single-Channel Speech Enhancement“, *In Proc. of the 42nd IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, S. 4980–4984, Mar. 2017.

[CHUTM13] A. Chinaev, R. Haeb-Umbach, J. Taghia und R. Martin: „Improved Single-Channel Nonstationary Noise Tracking by an Optimized MAP-Based Postprocessor“, *In Proc. of the 38th Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, S. 7477–7481, May 2013.

[CKTVHU12] A. Chinaev, A. Krueger, D. H. Tran-Vu und R. Haeb-Umbach: „Improved Noise Power Spectral Density Tracking by a MAP-Based Postprocessor“,

In Proc. of the 37th Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), S. 4041–4044, Mar. 2012.

[CPHU14] A. Chinaev, M. Puels und R. Haeb-Umbach: „Spectral Noise Tracking for Improved Nonstationary Noise Robust ASR“, *In Proc. of the 11th ITG Symposium on Speech Communication*, S. 1–4, Sept. 2014.

[DCTHU14a] L. Drude, A. Chinaev, D. H. Tran Vu und R. Haeb-Umbach: „Source Counting in Speech Mixtures Using a Variational EM Approach for Complex Watson Mixture Models“, *In Proc. the of 39th Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, S. 6834–6838, May 2014.

[DCTHU14b] L. Drude, A. Chinaev, D. H. Tran Vu und R. Haeb-Umbach: „Towards Online Source Counting in Speech Mixtures Applying a Variational EM for Complex Watson Mixture Models“, *In Proc. of the 14th Int. Workshop on Acoustic Signal Enhancement (IWAENC)*, S. 213–217, Sept. 2014.

[HDCHU15] J. Heymann, L. Drude, A. Chinaev und R. Haeb-Umbach: „BLSTM supported GEV Beamformer Front-End for the 3RD CHiME Challenge“, *In Proc. of Automatic Speech Recognition and Understanding Workshop (ASRU)*, Dec. 2015.

	ICASSP	ITG	INTERSPEECH	ASRU	IWAENC
2012	[CKTVHU12]*	[CHU12]*	-	-	-
2013	[CHUTM13]* [CHU13]	-	-	-	-
2014	[DCTHU14a]	[CPHU14]*	-	-	[DCTHU14b]
2015	-	-	[CHU15]*	[HDCHU15]	-
2016	-	[CHHU16]* [CHDHU16]*	[CHU16]*	-	-
2017	[CHU17]*	-	-	-	-

Table A1.: Übersicht über Publikationen mit eigener Beteiligung. In Teilen wurden bestimmte Inhalte dieser Arbeit auch in den Publikationen vorgestellt, die mit Sternchen markiert sind.