Lehrstuhl für Mensch-Maschine-Kommunikation
Technische Universität München

Robuste Spracherkennung auf der Basis recheneffizienter auditiver Modelle

Ronald Römer

Vollständiger Abdruck der von der Fakultät für Elektrotechnik und Informationstechnik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktor-Ingenieurs (Dr.-Ing.)

genehmigten Dissertation.

Vorsitzender:		Univ.-Prof. Dr.sc.techn. A. Herkersdorf
Prüfer der Dissertation:	1.	apl. Prof. Dr.-Ing., Dr.-Ing. habil. G. Ruske
	2.	Univ.-Prof. Dr.-Ing. habil. R. Hoffmann, Technische Universität Dresden

Die Dissertation wurde am 04.09.2008 bei der Technischen Universität München eingereicht und durch die Fakultät für Elektrotechnik und Informationstechnik am 26.01.2009 angenommen.

Vorwort

Betrachtet man die Forschungsaktivitäten zur Erhöhung der Robustheit, so kann man drei Hauptströmungen beobachten. Zum einen findet man den Zweig der akustischen Sprachverarbeitung, hier wird an Methoden zur Geräuschunterdrückung, Geräuschkompensation und Kanalkompensation gearbeitet. Zu diesem Zweig können darüber hinaus noch die mehrkanaligen Mikrofonsysteme gezählt werden. Auch in der zweiten Hauptströmung, der Akustischen Modellierung (z.B.: *Parallel Model Combination*) sind ebenfalls sehr erfolgreiche Methoden zur Erhöhung der Robustheit gefunden worden. Diese basieren vor allem auf den Prinzipien des maschinellen Lernens.

Der dritte Zweig, für den stellvertretend die *Computational Auditory Scene Analysis* hervorgehoben werden sollte, beschäftigt sich stärker mit der Nachbildung von höheren Informationsverarbeitungsebenen, also der Erkennung, Identifikation und Separierung von verschiedenen akustischen Schallquellen. Dabei tritt das Problem der begrenzten Ressourcen (Echtzeitfähigkeit, Speicherbedarf) besonders stark in Erscheinung.

Die vorliegende Arbeit entstand daher vor dem Hintergrund, für die wesentlichen auditiven Wirkprinzipien vom Gehörgang bis hin zum Zentralen Auditorischen System geeignete echtzeitfähige Algorithmen zu entwickeln, die auch in sogenannten „embedded“ Systemen zur Anwendung kommen können. Die Arbeit ist demnach thematisch eher in die dritte Hauptströmung einzuordnen. Dabei konnte auf eine ganze Reihe theoretischer Vorarbeiten zurückgegriffen werden, an dieser Stelle seien beispielhaft die Arbeiten von S. Shamma, S. Seneff und O. Ghitza hervorgehoben. Erst mit diesen Arbeiten konnte das notwendige Verständnis für die Modellierung der wichtigsten Kerneigenschaften des Auditorischen Systems erworben werden. Gleiches gilt für einige Implementierungsvorschläge, hier sei insbesondere F. Perdigao genannt.

In der Natur lässt sich immer wieder das Phänomen beobachten, dass wenn sich bestimmte Prinzipien für ein biologisches System als wirksam erwiesen haben, diese dann häufig auch für andere Aufgaben Verwendung finden. Da nach derzeitigem Kenntnisstand die Informationsübertragung bei allen Sinnen ausschließlich unter Verwendung von Spikefolgen erfolgt, könnte ein solches Ökonomieprinzip auch für neuronale Signalverarbeitungsketten existieren. Dieses würde dann nicht nur in den verschiedenen Verarbeitungsebenen eines Sinnes sondern auch bei den anderen Sinnen zu beobachten sein. Ein solches Prinzip für neuronale Verarbeitungsketten wurde bereits 1978 von V. Mountcastle zur Diskussion gestellt.

In dieser Arbeit wird gezeigt, dass sich die recheneffiziente Umsetzung biologischer Wirkprinzipien auch in kleineren automatischen Spracherkennungssystemen – und dort insbesondere zur Erhöhung der Robustheit gegenüber Hintergrundstörungen – gewinnbringend integrieren lässt. Unter Berufung auf das Ökonomieprinzip können solche Wirkprinzipien in einem hierarchisch strukturierten Kommunikationssystem mehrfach zur Anwendung kommen. Dass solche Ansätze ein großes Potential für die höheren Ebenen der Mensch-Maschine-Kommunikation aufweisen, wird im Ausblick auf zukünftige Arbeiten angedeutet.

Vor diesem Hintergrund bin ich meinem Doktorvater Herrn Prof. Dr.-Ing. habil. G. Ruske zu besonderem Dank verpflichtet, er hat mir den größtmöglichen Freiraum für diese Arbeit zugesichert und auch in schwierigen Zeiten an der Betreuung festgehalten. Ebenso möchte ich mich bei Herrn Prof. Dr.-Ing. habil. R. Hoffmann für das Interesse an meiner Arbeit und für die Übernahme des Zweitgutachtens bedanken. Bedanken möchte ich mich auch bei meinen Arbeitskollegen Herrn Dipl.-Ing. R. Brückner und Herrn Prof. Dr. G. Wirsching für ihre zahlreichen Diskussionen, Anregungen und ihre stete Hilfsbereitschaft.

Nicht zuletzt möchte ich mich bei meiner Frau N. Römer bedanken, sie hat mir den nötigen Rückhalt geboten, um diese Arbeit neben der beruflichen Belastung beenden zu können. Schließlich möchte ich diese Arbeit meiner verstorbenen Mutter M. Römer widmen, sie konnte die Fertigstellung der Arbeit nach kurzer schwerer Krankheit nicht mehr erleben.

Liste der verwendeten Abkürzungen

AKF	*Autokorrelationsfunktion*
ASR	*Automatic Speech Recognition*
AGC	*Automatic Gain Control*
BPA	*Belief Propagation Algorithm*
CAS	*Central Auditory System*
CASA	*Computational Auditory Scene Analysis*
CC	*Center Clipper*
CPT	*Conditional Probability Table*
CVC	*Consonant Vowel Consonant*
CWT	*Continuos Wavelet Transformation*
DCT	*Discrete Cosine Transformation*
DRT	*Diagnostic Rhyme Test*
DWT	*Discrete Wavelet Transformation*
EIH	*Ensemble Interval Histogram*
ERB	*Equivalent Rectangular Bandwidth*
FFT	*Fast Fourier Transformation*
FWT	*Fast Wavelet Transformation*
HC	*Haircell*
HMM	*Hidden Markov Model*
HSR	*Human Speech Recognition*
HWR	*Half Wave Rectification*
IFFT	*Inverse Fast Fourier Transformation*
LDA	*Lineare Diskriminanzanalyse*
LIN	*Lateral Inhibition Network*
LPC	*Linear Predictive Coding*
LPCC	*Linear Predictive Cepstral Coefficients*
MFCC	*Mel Frequency Cepstral Coefficients*
PCF	*Pitch Coherent Features*
PAS	*Periphere Auditory System*
PLP	*Perceptual Linear Predictive*
RMS	*Root Mean Square*
RASTA	*Relative Spectra*
SBCOR	*Subband Correlation*

SM	*Subtraction Method*
SNR	*Signal Noise Ratio*
STFT	*Short Time Fourier Transformation*
TEO	*Teager Energy Operator*
TNC	*Time Normalized Correlogram*
WA	*Wortakkuratheit*
WT	*Wavelet Transformation*
VVM	*Vorverarbeitungsmethode*
VAMIG	*Vereinheitlichtes Auditives Modell mit Integrierter Geräuschunterdrückung*

Inhaltsverzeichnis

1 Einführung

Die einer Sprache zugrunde liegende Vielfalt und Variabilität – man denke nur an die verschiedenen Dialekte einer Sprache, die unterschiedliche Aussprache verschiedener Sprecher, deren Sprechgeschwindigkeiten oder auch an die Anzahl unterscheidbarer Worte – machen die automatische Spracherkennung zu einer anspruchsvollen Aufgabe an deren Lösung verschiedene wissenschaftliche Disziplinen beteiligt sind.

Mit der Verfügbarkeit großer Sprachstichproben und dem Einzug statistischer Verfahren in der Automatischen Spracherkennung (*ASR*), konnte die oben angesprochene Vielfalt und Variabilität in ausreichendem Maße berücksichtigt werden. Insbesondere mit den sogenannten ***H**idden **M**arkov **M**odellen* ist es möglich, akustische Ereignisse robust darzustellen und zu beschreiben. Der in diesem Zusammenhang verwendete Begriff der Robustheit bezieht sich zunächst ausschließlich auf die natürliche Variabilität und Vielfalt von Sprache. Anspruchsvolle praktische Anwendungen fordern darüber hinaus auch eine zuverlässige Funktionsweise des Erkennungsprozesses bei gestörter Umgebung. Naturgemäß fällt dabei der Repräsentation von akustischen Ereignissen eine große Bedeutung zu. Deren Aufgabe ist es, die Sprachdaten von irrelevanten oder störenden Anteilen zu trennen und in eine möglichst kompakte Merkmalsdarstellung zu überführen. Dies schließt in der Regel verschiedene Vorverarbeitungsschritte ein.

Geeignete Merkmale können bspw. mit den Methoden der Signalanalyse (*DFT, LPC*, Wavelet- Zerlegung usw.) gefunden werden. Das Ziel einer solchen Analyse besteht darin, diejenigen Merkmale des Sprachsignals zu extrahieren, welche die relevante Information über das Sprachsignal tragen. Einen weiteren Aspekt der Merkmalsdarstellung stellt die Kompaktheit der Merkmale dar, d.h. aus ökonomischer Sicht sollten Merkmale aus möglichst wenigen voneinander unabhängigen Komponenten bestehen. Die gesamte Verarbeitungskette vom Eingangssignal über die Rauschunterdrückung bis hin zur Bildung einer Folge von Merkmalvektoren wird im Folgenden als ***V**or-**V**erarbeitungs-**M**ethode* (*VVM)* bezeichnet.

Aus historischer Sicht wurde zur Analyse von Sprachsignalen vorzugsweise Sprachmaterial aus geräuscharmer Umgebung herangezogen. Auf Analysen mit derartigem Sprachmaterial basieren die derzeit am häufigsten verwendeten Merkmale wie *MFCC* oder *LPCC*. Verwendet man diese Merkmale in geräuschbehafteten Szenarien ohne den Einfluss der Störungen durch geeignete Kompensations- oder Unterdrückungsmethoden zu reduzieren, verschlechtert sich die Qualität eines *ASR*-Systems deutlich. Im Gegensatz zu den *ASR*-Systemen zeigen Untersuchungen des menschlichen auditiven Systems (*HSR*) deren Überlegenheit über einen weiten Bereich verschiedener Störszenarien *[Lippmann-97]*. Daraus lässt sich schlussfolgern, dass das Auditive System - bedingt durch evolutionären Druck - derart optimiert wurde, dass es über einen weiten Bereich akustischer Störungen bei unterschiedlichen Störintensitäten zuverlässig arbeitet. Das bedeutet aber, dass die wichtigsten Informationen in einem Sprachsignal durch diejenigen Eigenschaften des Sprachsignals repräsentiert werden, welche vom *HSR* über einen weiten Bereich von Störungen zuverlässig geschätzt werden können.

An dieser Stelle scheint ein Paradigmenwechsel möglich: Anstelle der Suche nach einer optimalen Merkmalsdarstellung basierend auf Sprachdaten in geräuscharmer Umgebung, sollte nach denjenigen Merkmalen gesucht werden, welche in einem weiten Bereich verschiedener Störszenarien mit ausreichender Sicherheit geschätzt werden können. Ein angemessenes Konzept zur Lösung des Problems stellt die Einbeziehung der Zuverlässigkeit oder Robustheit der extrahierten Merkmale dar.

Um eine gewisse Einschränkung hinsichtlich der in Frage kommenden Merkmale zu gewinnen, muss zunächst die Struktur von Sprachsignalen analysiert werden. In der Literatur werden zu diesem Zweck die Mechanismen von Sprachproduktion und Sprachwahrnehmung durch entsprechende Modelle und deren Parametrisierung beschrieben. Für die Sprachproduktion wird häufig das Quellen-Filter-Modell nach angegeben *[Fant-60]*, *[Flanagan-72]*.

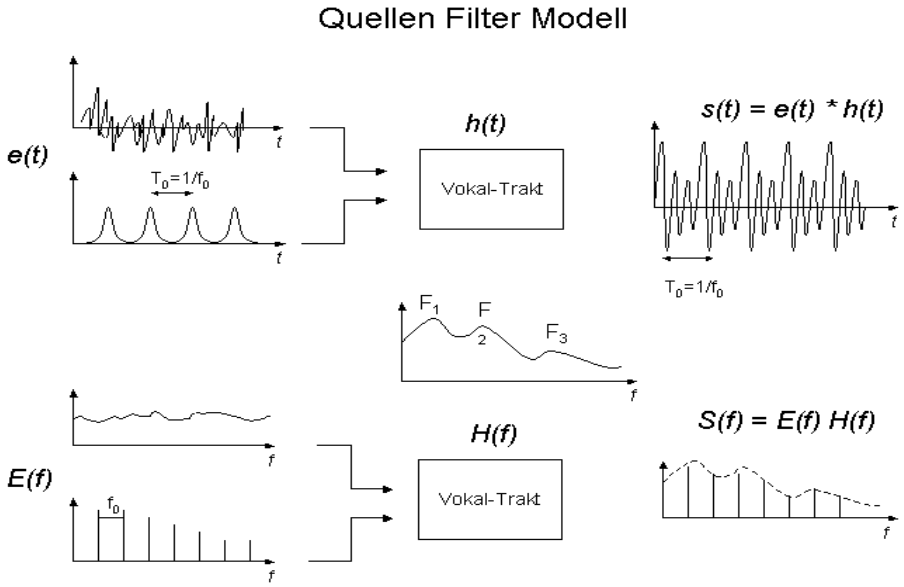

Abbildung 1.1 ***Modell der Sprachproduktion. Der obere Teil der Abbildung zeigt die Zusammenhänge im Zeitbereich. In diesem Modell wird die Erzeugung von stimmhaften Lauten durch die Erregung des Vokaltraktes mit einer quasiperiodischen Pulsfolge erklärt. Bei stimmlosen Lauten tritt an die Stelle der quasiperiodischen Anregung eine regellose aperiodische Erregung. Es sind auch Überlagerungen beider Anregungsarten möglich. Die Resonanzstellen des Vokaltraktes führen im Spektrum zu energiereichen Spitzen, sie werden Formanten genannt. Mathematische kann das Sprachsignal durch eine Faltung des erregenden Signals mit der Impulsantwort des Vokaltraktes beschrieben werden. Der untere Teil der Abbildung beschreibt die Zusammenhänge im Frequenzbereich als Multiplikation von Anregungsspektrum und Frequenzantwort des Vokaltraktes.***

Nach Abbildung 1.1 wird zwischen stimmhafter und stimmloser Anregung unterschieden. Die spektrale Formung des Sprachsignals wird durch die Resonanzfrequenzen bzw. die Formanten des Vokaltraktes bestimmt. Bei stimmhafter also periodischer Erregung findet man ein diskretes Sprachspektrum vor, bei dem die Frequenzantwort des Vokaltraktes durch Vielfache der Grundfrequenz F_0 abgetastet wird. Im stimmlosen Fall dagegen, ist die Erregung von aperiodischer Natur und hat ein kontinuierliches Spektrum. Das Sprachspektrum wird nach dem Quelle-Filter-Modell als das Produkt von Frequenzantwort des Vokaltraktes und des erregenden Spektrums verstanden. Die unterschiedlichen Stellungen der Artikulationsorgane führen zu Änderungen in der Geometrie des Vokaltraktes und somit zu entsprechenden Modifikation der Frequenzantwort des Vokaltraktes. In den meisten ASR-Systemen liegt der Schwerpunkt bei der Merkmalsbestimmung im Kern auf der Schätzung der Frequenzantwort des Vokaltraktes, da die zeitliche Entwicklung der Formanten als wesentliche Informationsquelle bezüglich der Identität von Sprachlauten betrachtet wird. Die Grundfrequenz dagegen wird häufig zur Unterscheidung des Erregungstyps oder einer Geschlechtsklassifikation verwendet.

Bezüglich der Sprachwahrnehmung dominiert das Filterbank-Modell. Das Ohr wird hier als eine Art Frequenzanalysator verstanden, wobei die einzelnen Frequenzen des akustischen Signals auf bestimmte Orte der Basilarmembran abgebildet werden und an dieser Stelle die so genannten Haarzellen erregen. Die Haarzellen wiederum erzeugen aus den entsprechenden Erregungsmustern auf der Basilarmembran neuronale Aktivitäten, welche dann auf den Nervenbahnen zum Gehirn geleitet werden. Neuere Untersuchungen zeigen, dass auch die zeitliche Struktur der Einhüllenden der Teilbandsignale der Filterbank im Gehirn abgebildet wird und diese Information auch zur Auswertung zur Verfügung steht. So konnte von *[Dau-99]* beobachtet werden, dass die einem Träger aufgeprägten Amplitudenmodulationen ebenfalls auf bestimmte Gehirnareale abgebildet werden.

Für eine robuste Sprachkommunikation sollten allerdings weitere Aspekte berücksichtigt werden. Zu diesem Zweck wird eine vollständige aber stark vereinfachte Kommunikationskette – von der Idee einer Nachricht beim Sender, über die Mechanismen der Sprachproduktion und der Sprachwahrnehmung bis hin zur Manifestation der Nachricht beim Empfänger – betrachtet, dabei kann die Kommunikationskette durch drei verschiedene Abstraktionsebenen charakterisiert werden: linguistische-, physiologische- und akustische Ebene. Insbesondere die physiologische Ebene kann feiner unterschieden werden: So gelangt man neben der Strukturanalyse von Sprachsignalen auch zu Anhaltspunkten bezüglich möglicher Strategien zur Gewinnung von robusten Merkmalen.

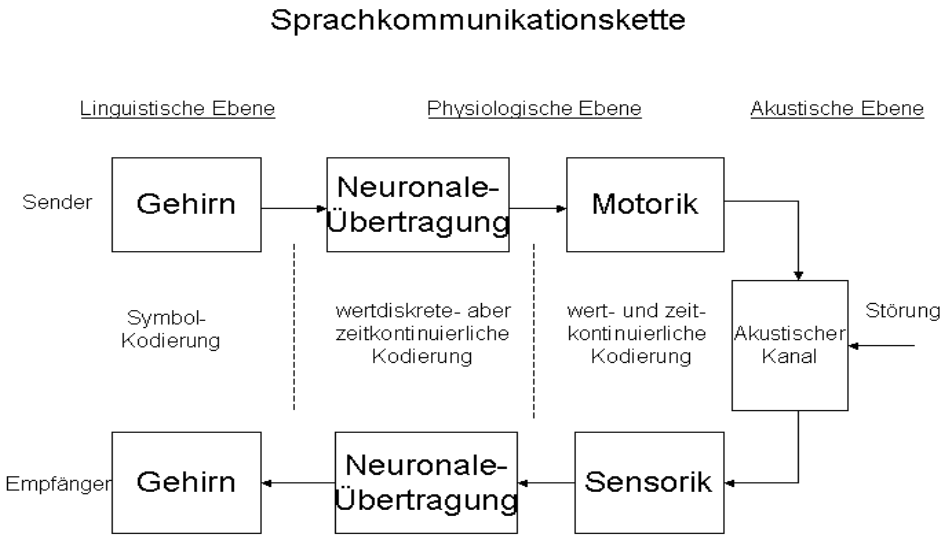

Abbildung 1.2 ***Die Sprachkommunikationskette kann hinsichtlich der Nachrichtenübertragung durch drei Abstraktionsebenen charakterisiert werden. Betrachtet man diese Kette aus der Sicht der Signalkodierung, so kann ein Teil der physiologischen Ebene als Bindeglied zur akustischen Ebene aufgefasst werden. Die Signalübertragung von Motorik, Sensorik und der akustische Ebene ist durch eine wert- und zeitkontinuierliche Kodierung gekennzeichnet. Diese Form der Signalkodierung scheint zur Extraktion zuverlässiger Sprachmerkmale in besonderem Maße geeignet zu sein, da Redundanz und Kontinuität hier die bestimmenden Eigenschaften des Signals sind. Dagegen liegen die neuronalen Signale in einer zeitkontinuierlichen aber wertdiskreten Kodierung vor. Wahrscheinlich ist diese Kodierungsform einerseits für die Übertragung und Speicherung von verschiedenen Hypothesen einer Nachricht geeignet und stellt andererseits eine optimale Basis für Entscheidungsprozesse dar. Für die letzte Behauptung sprechen Beobachtungen, dass die zeitlichen Abstände einzelner neuronaler Spikes in der Größenordnung von handlungsrelevanter Zeiten (etwa 5...40 ms) liegen.***

Eine zu übermittelnde Nachricht liegt zunächst im Gehirn des sendenden Individuums in Form einer abstrakten Idee vor und wird auf der neuronalen Ebene durch Spikefolgen übertragen. Nach derzeitigem Kenntnisstand wird die zu übertragende Information auf dieser Ebene vor allem durch die zeitliche Abfolge der einzelnen Spikes kodiert. Im Prinzip handelt es sich hier um eine zeitkontinuierliche aber wertediskrete Übertragung der Nachricht, welche aus nachrichtentechnischer Sicht eine gewisse Ähnlichkeit zur Pulsphasen- bzw. Pulspositionsmodulation zeigt. Die Spikefolgen wiederum steuern die neuronalen Elemente der Sprachmotorik (Artikulationsorgane, Aktivierung der Erregungsenergie usw.), dabei ist zu beachten, dass die Bewegung der Artikulationsorgane wieder wertkontinuierlich erfolgt. Das derart erzeugte Sprachsignal wird anschließend akustisch abgestrahlt. Nunmehr dienen akustische Wellen als Träger der Sprachinformation und übertragen diese zum Empfänger. Auf diesem Wege können sich dem Sprachsignal zusätzliche Signale anderer Quellen überlagern. Außerdem kann der spezifische Übertragungsweg zu Situationen führen, in denen das Signal Reflexionen, Auslöschungen, Verstärkungen usw. unterworfen wird.

Am Empfangsort liegt daher ein komplexes akustisches Signal vor, welches aus einer Vielzahl von zeitabhängigen Störungen resultiert. Die Empfangsorgane verarbeiten die akustische Information zunächst auch wieder auf kontinuierlichem Weg, ehe das empfangene akustische Signal nach Erregung der Haarzellen wiederum in Spikefolgen kodiert wird. Über die entsprechenden Nervenbahnen werden die Spikefolgen zum Gehirn des Empfängers geleitet und in bisher noch unbekannter Weise in sprachliche Symbole transformiert. Unter Verwendung von grammatikalischem Wissen kann das Gehirn aus der dekodierten Symbolfolge auf die dazugehörige Idee schließen, so dass sich nun im Bewusstsein des Empfängers eine mögliche Nachricht manifestieren kann.

Mit dieser Sicht auf die Kommunikationskette kann man feststellen, dass die zu übertragende Information zunächst mindestens zwei unterschiedliche Signalkodierungen durchläuft: Zeitkontinuierlich und wertediskret auf physiologischer Ebene, sowie zeit- und wertekontinuierlich auf der akustischen Ebene. Bei diesen beiden Ebenen ist es sicher gerechtfertigt von Signalverarbeitung oder subsymbolischer Informationsverarbeitung zu sprechen, dagegen muss man sich auf der linguistischen Ebene mit symbolischer Informationsverarbeitung befassen. In diesen Bereich fallen bspw. Sprachmodelle und Grammatik. Die hier vorliegende Arbeit beschäftigt sich jedoch nur mit den ersten beiden Ebenen, höheres Wissen der linguistischen Ebene wird hier ausgeschlossen.

Nun ist einerseits bekannt, dass das Sprachsignal auf der akustischen Ebene hochgradig redundant ist, und eben dort auch die oben angesprochenen Störszenarien wirksam werden. Andererseits ist von der neuronalen Ebene ebenso bekannt *[Rieke-99]*, dass sich das zeitliche Auftreten einzelner Spikes in der Größenordnung von handlungsrelevanten Zeiten bewegt. Aus evolutionärer Sicht kann man nun daraus schließen, dass den einzelnen Spikes ein zuverlässiger Auslöseprozess auf der Basis robuster Merkmale vorangegangen sein sollte und eine derart generierte Spikefolge nun weit weniger Redundanz enthält. Hierfür findet man in eine Reihe von Indizien:

Bereits in den 1960 Jahren postulierten die Autoren in *[Barlow-61]*, dass Nervenzellen auf einen Reiz mit einem möglichst geringen Aufwand antworten sollten, also mit so wenig Redundanz wie möglich. Tatsächlich gibt es experimentelle Hinweise dafür, dass die Kodierung von Reizen durch sensorische Neuronen kaum redundant ist. Dabei wird überprüft, ob die Zuverlässigkeit von Entscheidungen auf der Wahrnehmungsebene mit der Zuverlässigkeit von Entscheidungen vergleichbar ist, welche auf der Basis der Antwort eines einzelnen Neurons erfolgen.

Verschiedene Experimente bezüglich der Trennung von verschiedenen Reizklassen haben nun gezeigt, dass sowohl im auditiven als auch im visuellen System die Wahrscheinlichkeit korrekter Entscheidungen basierend auf den Reizantworten einzelner Neuronen kaum von der Wahrscheinlichkeit von korrekten Entscheidungen auf der Verhaltens- bzw. Wahrnehmungsebene abweicht. Diese Beobachtung ändert sich auch dann nicht wesentlich, wenn die Antworten benachbarter Neuronen zur Entscheidung herangezogen werden. Ein häufig angeführtes Argument gegen die effiziente und redundanzarme Kodierung von Neuronen ist die mäßige Reproduzierbarkeit (begründet durch die Varianz im zeitlichen Auftreten von Spikes) neuronaler Antworten. Dieses Argument ist eng mit dem Streit darüber verbunden, ob Neuronen einen Raten- oder Timingkode verwenden.

Dieser Diskurs konnte durch eine Reihe von Experimenten *[Rieke-99]* beendet werden, bei denen statt statischer Reize nunmehr natürliche Stimuli verwendet wurden. Hier zeigte es sich, dass die zugehörige Spikefolge zuverlässig reproduziert werden kann, wenn das stimulierende Signal innere Korrelationen aufweißt. Aus diesem Grund verwenden Neuronen wahrscheinlich beide Kodierungsformen. Wenn die Änderungsgeschwindigkeit interner Korrelationen niedrig genug ist wird der Timingkode verwendet, bei statischen Stimuli oder bei Stimuli mit sehr schnell veränderlichen internen Korrelationen wird der Ratenkode verwendet. Darüber hinaus haben die Autoren gezeigt, dass sich das stimulierende Signal aus der zugehörigen neuronalen (dünn besetzten) Spikefolge rekonstruieren lässt. Aus der Qualität der Rekonstruktion konnten die Forscher sogar rechnerisch abschätzen, *wie viel* Information das Neuron übertragen haben musste – je geringer der Rekonstruktionsfehler, desto mehr Information.

Demnach können Neuronen die eingehenden Signale sogar nahe der physikalischen Grenze kodieren, wenn die Signale die entsprechenden internen Korrelationen aufweisen. Bekanntermaßen gilt dies für den überwiegenden Teil der auf natürliche Art und Weise erzeugten Signale, insbesondere aber eben auch für die Klasse der Sprachsignale. Nach den oben angeführten Argumenten wäre es demnach durchaus plausibel, die Bestimmung möglichst zuverlässiger Merkmale auf der wert- und zeitkontinuierlichen Ebene durchzuführen. Auf dieser Ebene ist die nötige Redundanz und Kontinuität im Signal tatsächlich noch vorhanden.

Die Überführung wert- und zeitkontinuierlicher Signale in zeitkontinuierliche- aber wertdiskrete Signale wird im Peripheren Auditiven System von den Haarzellen übernommen. Diese stellen das Bindeglied zur neuronalen Informationsverarbeitung dar und sollten demnach einen wichtigen Beitrag zur Zuverlässigkeit von Entscheidungen leisten. D.h. es spricht einiges dafür, das Übertragungsverhalten von Haarzellen zu modellieren und unter Störeinflüssen zu analysieren.

Andere Forschungsansätze konzentrieren sich nicht vordergründig auf die Nachbildung von auditorischen Wirkprinzipien sondern verwenden andere Werkzeuge um die Struktur des Sprachsignals zu analysieren, oftmals lassen sie sich dabei aber ebenfalls vom Redundanz- und Kontinuitätsprinzips leiten. Insbesondere vor dem Hintergrund der Vielzahl möglicher Störszenarien erfordert eine effiziente Sprachkommunikation, dass die zeitliche Änderung von Sprachsignalkomponenten einerseits langsam genug ist, um auch im Rauschen zuverlässig detektiert zu werden und anderseits eine genügende schnelle Änderung dieser Komponenten, um genügend Information in handlungsrelevanten Zeiten zu übertragen.

Gemäß diesem Prinzip werden in *[Andringa-02]* Methoden zur Merkmalsextraktion beschrieben, die auf dem Einsatz einer linearen Filterbank gefolgt von verschiedenen Korrelationsanalysen (***T****ime* ***N****ormalized* ***C****orrelations*) zur Extraktion charakteristischer Periodizitäten sowohl innerhalb als auch zwischen den einzelnen Teilbändern beruhen. Die *TNC*- Methode ist zwar ein sehr mächtiges Werkzeug mit dem die innere Struktur des Sprachsignals erfasst werden kann, es enthält aber nur in sehr geringem Umfang Elemente des peripheren auditiven Systems und ist leider durch einen besonders hohen Rechenaufwand charakterisiert. Dies folgt aber unmittelbar als Konsequenz aus dem Redundanz- und Kontinuitätsprinzip. Andererseits ermöglichen solche hochauflösenden Verfahren überhaupt erst die genaue Identifikation und Beobachtung von sich langsam entwickelnden Signalkomponenten, welche gleichzeitig einsetzen und sich anschließend relativ langsam und kohärent zueinander entwickeln, wie das z.B. bei den Harmonischen der Sprachgrundfrequenz der Fall ist. Dass darüber hinaus die Sprachgrundfrequenz und deren Harmonische auch wirksam zur Merkmalsextraktion in stimmhaften Abschnitten herangezogen werden kann, ist daher ein besonderer Verdienst einer solchen Arbeit, auch wenn noch keine echtzeitfähige Lösung präsentiert werden konnte.

Neben der Nachbildung von auditiven Eigenschaften des Peripheren Auditiven Systems sind in jüngerer Zeit verstärkte Anstrengungen unternommen worden, um auch Eigenschaften des Zentralen Auditiven Systems modellieren zu können. Eine dieser Eigenschaften ist die sogenannte Bindungseigenschaft. Vom visuellen System ist bspw. bekannt, dass synchron feuernde Spikes zusammengehörige Merkmale eines einzigen wahrgenommenen Objekts kodieren und somit der Segmentierung von unterschiedlichen visuellen Objekten innerhalb eines Bildes dienen. Eine Vielzahl von Forschern vertritt heute die Meinung, dass unser Gehirn grundlegende Prinzipien der neuronalen Informationsverarbeitung unabhängig von den spezifischen Sinneskanälen verwendet. D.h. es kann vermutetet werden, dass das Prinzip der Merkmalsbindung auch im auditiven System zur Anwendung kommt. Demnach sollten zusammengehörige Signalkomponenten eines auditiven Objekts hervorgehoben und vom akustischen Hintergrund getrennt werden können. Eine wichtige Voraussetzung zur Bildung von auditiven Objekten besteht zuallererst darin, zusammengehörige Signalkomponenten identifizieren zu können. Ein mögliches Prinzip ist mit der Identifikation der Harmonischen der Sprachgrundfrequenz bereits genannt worden.

Ein bisher in Sprachverarbeitungssystemen wenig beachteter struktureller Aspekt des auditiven Systems bildet die Grundlage einer weiteren zu modellierenden Eigenschaft des Zentralen Auditiven Systems. So ist bspw. bekannt, dass der Informationsfluss von den Sinneszellen über das Periphere Auditorische System bis hin zum Zentralen Auditorischen Systems sowohl auf afferenten (aufsteigenden) als auch auf efferenten (absteigenden) Nervenbahnen stattfindet.

Zudem zeigen Untersuchungen in der Hirnforschung, dass ein inneres Abbild der motorischen Strukturen und deren Dynamik ausgenutzt werden kann, um Eingangssignale vorherzusagen. Nun kann man vermuten, dass auf den efferenten Bahnen Vorhersagen bezüglich des zu erwartenden Eingangssignals zurückgeführt werden und dass diese Vorhersagen in stimmhaften Abschnitten stärker berücksichtigt werden als in stimmlosen Abschnitten. D.h. derartige Vorhersagen können bspw. genutzt werden, um Mehrdeutigkeiten aufzulösen.

Zusammenfassung

Die vorliegende Arbeit zielt in ihrem ersten Teil auf eine Modellierung ab, bei der die Bestimmung zuverlässiger Merkmale auf der konsequenten Verwendung von auditiven Wirkprinzipien beruht. Dabei sollen sowohl die Elemente des Peripheren Auditiven Systems wie gehörgerechte Filterbank, Haarzellenmodell und Laterale Inhibition als auch Wirkprinzipien des Zentralen Auditiven Systems wie Bindungseigenschaft und die Einbeziehung der Vorhersage in stimmhaften Abschnitten berücksichtigt werden. Um nun die verschiedenen Realisierungsmöglichkeiten Auditiver Modelle miteinander vergleichen zu können, wird das ***V**ereinheitlichte **A**uditive **M**odell mit **I**ntegrierter **G**eräuschunterdrückung* auf der Grundlage der Artikulationstheorie eingeführt. Das *VAMIG* zeichnet sich vor allem dadurch aus, dass jedes *VAMIG*-Modul in unterschiedlichen Ausprägungen realisiert werden kann. Somit kann ein breites Spektrum unterschiedlicher Realisierungsmöglichkeiten, angefangen vom klassischen *MFCC*-Modell über verschiedene hybride Ausprägungen bis hin zum reinen auditiven Modell, untersucht werden.

Die Artikulationstheorie wurde bereits 1953 von H. Fletcher veröffentlicht, sie hatte ihren Ursprung in einer Sprachsignal-Analysemethode, bei der man zur Bestimmung der Verständlichkeit von gestörter Telefonsprache die Einbeziehung von Kontextwissen vermeiden musste. Das aus den Experimenten resultierende Multiband- Modell legte einige grundlegende Prinzipien der robusten Merkmalsbestimmung des *HSR*-Systems offen.

Mit der experimentellen Bestimmung der Robustheit der verschiedenen *VAMIG*-Ausprägungen und den dazugehörigen Erkennexperimenten beschäftigt sich der zweite Teil der Arbeit. Mit den sich ergebenden Schlussfolgerungen und einem Ausblick auf weitere Maßnahmen zur Optimierung der hier vorgestellten Auditiven Modelle wird diese Arbeit zunächst abgeschlossen. Weitere über diese Arbeit hinausgehende Vorschläge deuten schließlich an, wie biologische motivierte Wirkprinzipien gemäß dem Ökonomieprinzip auch in den höheren Sprachverarbeitungsebenen konsequent angewendet werden können.

Zur Gliederung der vorliegenden Arbeit

Die Dissertation gliedert sich in 10 Kapitel. Nach den einführenden Worten wird zunächst ein kurzer Überblick zum Inhalt der nächsten Kapitel gegeben. Im Anhang A werden die wichtigsten statistischen Grundlagen zusammengefasst. Die Kernaussagen der Artikulationstheorie *[Allen-94]* werden wegen deren besonderen Bedeutung für diese Arbeit in Anhang B herausgearbeitet. Schließlich erfolgt in Anhang C eine Zusammenfassung der wichtigsten mathematischen Zusammenhänge, welche zum Verständnis der Wavelet-Transformation beitragen können.

Kapitel 2

In Kapitel 2 wird die Artikulationstheorie von H. Fletcher als motivierendes Element aber auch als Gerüst dieser Arbeit vorgestellt. Zunächst sollen die Gründe hierfür dargelegt werden. In *[Allen-94]* wurde eine Rückschau auf die Arbeiten von H. Fletcher – vor allem dessen Artikulationstheorie – im Lichte des heutigen Wissensstandes gegeben. Seitdem gab es eine Vielzahl von Veröffentlichungen, welche sich mit einigen Teilaspekten der Artikulationstheorie beschäftigten und die zu neuen Erkenntnissen und Anregungen auf dem Gebiet der automatischen Spracherkennung führten *[Hermansky-99]*, *[Mirghafori-99]*, *[Bourlard-96]*. Vergleicht man bspw. die heutigen Phonemerkennraten mit der menschlichen Performanz, so zeigen sich deutliche Nachteile gegenüber dem biologischen Vorbild. Dies gilt insbesondere dann, wenn die Erkennung in einem mit akustischen Störungen behafteten Umfeld stattfindet. Das lässt darauf schließen, dass wichtige Mechanismen in den aktuellen Modellen noch nicht berücksichtigt sind. Es liegt nahe, die oben genannten Mechanismen zunächst bei denjenigen Modellen zu suchen, welche in den Disziplinen Physiologie (Aufbau und Funktionsweise der Sprechorgane und des Gehörs) und Psychoakustik (Wahrnehmung von akustischen Signalen) ihre Wurzeln haben.

Kapitel 3

Gegenstand von Kapitel 3 ist die Identifikation derjenigen Mechanismen und Wirkprinzipien von auditiven Modellen, welche bei der Übertragung von linguistisch wichtigen Informationen das SNR erhöhen können. Nach einer kurzen Beschreibung des Aufbaus des Peripheren Auditiven Systems wird auf die relevanten statischen und dynamischen Eigenschaften auditiver Modelle eingegangen. Zu den statischen Eigenschaften gehören vor allem das Konzept der Lautheit und das Konzept der Frequenzgruppen. Die dynamischen Eigenschaften finden sich in verschiedenen Modulationseffekten, dem Phänomen der Adaption und den mit der Adaption eng verwandten Maskierungseffekten wieder. Ein weiteres Charakteristikum auditiver Modelle sind die sogenannten ***L**ateralen **I**nhibitions **N**etzwerke***,** derartige Netzwerke findet man häufig in der Verarbeitung von neuronalen Signalen. In dieser Arbeit dienen sie vor allem der Redundanzreduktion von benachbarten Hörkanälen. Messungen von Hör- Filterkurven haben gezeigt, dass sich die Frequenzbänder mit zunehmender Mittenfrequenz immer stärker überlappen. Die damit verbundene Zunahme der Filterbandbreiten geht einher mit einer immer feineren Zeitauflösung. Um die ebenfalls zunehmende Redundanz in den Frequenzbändern zu reduzieren, werden Signalanteile aus den Nachbarbändern unterdrückt. Dies führt letztlich auf voneinander unabhängige Teilbandsignale. Im Weiteren werden drei aus der Literatur bekannte Modellierungen vorgestellt, mit denen die eben genannten Eigenschaften teilweise nachgebildet werden.

Das Modell von *[Yang-92]* zeichnet sich durch eine vollständige mathematische Formulierung derjenigen Signaltransformationen aus, welche die frühen Verarbeitungsstufen des Gehörs – Analyse, Signalwandlung und Reduktionsstufe – charakterisieren. Neben der Einführung einer mehrfach auflösenden Transformation wird das Prinzip der Selbstnormalisierung dargestellt und ein *LIN* beschrieben, welches auf der Berechnung der dominanten Frequenz basiert.

Das ebenfalls 3-stufige Modell von *[Seneff-84]* konzentriert sich vor allem auf die Modellierung eines nichtlinearen Haarzellenmodells mit dem das Phänomen der Adaption nachgebildet werden kann. Das Konzept der dominanten Frequenz bzw. der dominanten Periodizität wird hier genutzt, um auch bei Aussteuerungen innerhalb des Sättigungsbereichs der Haarzelle Formanten bei niedrigen SNR auflösen zu können.

Als letztes Modellbeispiel des Peripheren Auditiven Systems wird das ***E**nsemble **I**nterval **H**istogram* Modell *[Githza-94]* vorgestellt, es besteht aus einer gehörrichtigen Filterbank mit einem nachfolgenden nichtlinearen Prozessor, welcher die Ausgänge der Filter in entsprechende neuronale Feuermuster umsetzt.

Obwohl das *EIH* ausschließlich Frequenz- und Intensitäts- Informationen zur Modellierung neuronaler Feueraktivität verwendet, zeigt es eine bemerkenswerte Robustheit gegenüber weißem Rauschen und sporadisch auftretenden Signalspitzen. Die dynamischen Eigenschaften der Haarzelle werden in diesem Modell nicht berücksichtigt.

Auf den in diesen Modellen beschriebenen Kerneigenschaften des *PAS* werden alle recheneffizienten Modelle von Kapitel 4 zurückzuführen sein. Die Beschreibung des Zentralen Auditiven Systems bildet den Abschluss von Kapitel 3. Da dieses System bei weitem nicht so gut verstanden ist wie das Periphere Auditive System, kann die Beschreibung nicht so detailliert erfolgen und muss wohl auch mit einigen Spekulationen auskommen. Die derzeit am weitesten verbreitete Theorie ist unter dem Namen *Auditory Scene Analysis* bekannt *[Bregman-90]*. Den Kern dieses Abschnitts bilden die beiden Hypothesen zur Merkmalsbindung und der Vorhersage erwarteter Eingangssignale zur Auflösung von Mehrdeutigkeiten. Die Integration dieser Eigenschaften in einem recheneffizienten auditiven Modell wird in Kapitel 5 beschrieben.

Kapitel 4

Zur Analyse bestimmter auditiver Phänomene müssen die vorgestellten Modelle in Kapitel 3 sowohl über eine hohe Zeitauflösung als auch über eine hohe Frequenzauflösung verfügen. Aufgrund des damit verbundenen enormen Rechzeitbedarfs sind diese Modelle dann jedoch nicht direkt in echtzeitfähigen *ASR*-Systemen einsetzbar.

In diesem Kapitel werden zunächst zwei Modelle vorgestellt, in denen neben den statischen Eigenschaften auch die dynamischen Eigenschaften des *PAS* berücksichtigt wurden und in denen bereits eine dem Haarzellenmodell angepasste Rauschunterdrückung integriert ist. Diese Anpassung ist notwendig, da die Haarzellen der Signalübertragung nur einen begrenzten dynamischen Bereich zur Verfügung stellen können. In dem in *[Vereecken-95]* vorgestellten Modell ist sowohl die Filterbank als auch das Haarzellenmodell im Zeitbereich angesiedelt. Da in jedem Teilband mit der vollen Abtastrate gerechnet werden muss, ist die Bestimmung von Merkmalen relativ rechenaufwendig.

Eine Realisierung des Modells im Frequenzbereich mit einem ähnlichen Ansatz zur Geräuschunterdrückung, allerdings mit deutlich reduziertem Rechenaufwand, wurde später von *[Perdigao-99]* vorgeschlagen. Die einzelnen Teilbänder werden hier lediglich mit der Framerate verarbeitet. Weiterhin konnte gezeigt werden, dass für ein synthetisches Signal die Anwendung des Haarzellenmodells zu einer lokalen Verbesserung des SNR führt. Dieses Modell erfährt in seiner ursprünglichen Form eine Erweiterung durch ein *LIN*, welches ebenfalls auf dem Prinzip der dominanten Frequenz basiert.

Als letzte Variante einer *PAS*-Modellierung wird anstelle einer *FFT*-Analyse eine ***W****avelet*-***T****ransformation* verwendet, auch hier wird ein *LIN* verwendet, welches auf dem Prinzip der dominanten Frequenz basiert.

Kapitel 5

Hier werden zwei Modellvorschläge unterbreitet, mit denen die Hypothesen zur Merkmalsbindung und der Vorhersage zur Auflösung von Mehrdeutigkeiten experimentell überprüft werden können. Dabei beziehen sich alle Ausführungen auf ein Modell im Frequenzbereich, da bisher nur für dieses Modell eine relativ einfache recheneffiziente Implementierung gefunden wurde. Es werden zwei Ansätze verfolgt. Der erste Ansatz bewertet die kohärenten Signalkomponenten in der Teilbandebene nach dem *LIN*, im zweiten Ansatz dagegen erfolgt die Bewertung kohärenter Signalkomponenten bereits in der Frequenzebene. In beiden Fällen erfolgt die Bestimmung zueinander kohärenter Signalkomponenten unter Verwendung der Sprachgrundfrequenz und deren Harmonischen.

Kapitel 6

Der dynamische Bereich einer Haarzelle wird durch den linearen Bereich einer sigmoidalen Kennlinie definiert. Außerhalb des linearen Bereichs dominiert das Sättigungsverhalten der Kennlinie die Signalübertragung. Signaländerungen in diesem Bereich können nicht mehr übertragen werden. Bei normaler Aussteuerung bewegt sich das Signal innerhalb des dynamischen Bereichs, Überlagerungen von Geräuschen führen dagegen zu einer Verschiebung des Arbeitspunktes in den Bereich der Sättigung. Aus diesem Grund sind Methoden notwendig, mit denen das Signal unabhängig von der Stärke der Störung innerhalb des dynamischen Bereichs verbleiben kann. Die entsprechenden Methoden zur Geräuschunterdrückung für *VVM* mit einem Haarzellenmodell werden in diesem Kapitel detailliert beschrieben.

Kapitel 7

In den Abschnitten 3 bis 6 werden theoretische auditive Modelle, deren Implementierungen und verschiedene Verfahren zur Rauschunterdrückung vorgestellt. Um nun eine systematische Untersuchung von auditiven Modellen zu ermöglichen, wird in Kapitel 7 das ***V****ereinheitlichte* ***A****uditive* ***M****odell mit* ***I****ntegrierter* ***G****eräuschunterdrückung* (*VAMIG*) eingeführt. Dieses Modell zeichnet sich zunächst durch die Berücksichtigung der wesentlichen *PAS*- und *CAS* Eigenschaften aus, diese können jedoch unterschiedlich realisiert sein. Durch Variation einiger weniger *VAMIG*-Module kann man – angefangen vom klassischen *MFCC*-Modell über hybride Modelle bis hin zu rein auditiven Modellen – eine Vielzahl möglicher *VVM* erzeugen. Diese verschiedenen Ausprägungen des *VAMIG* lassen sich nun bezüglich ihrer Robustheit und Erkenngenauigkeit systematisch miteinander vergleichen.

Kapitel 8

Der Begriff der Robustheit ist zwar ein weitläufig verwendeter aber doch auch ein sehr unscharfer Begriff. In diesem Kapitel werden nun zwei konkrete Definitionen der Robustheit herausgearbeitet, mit denen die quantitative Bewertung der Robustheit einer *VVM* ermöglicht wird. Während die erste Definition einer informationstheoretischen Deutung entspricht und vor allem theoretischen Wert hat, lässt sich die zweite Definition direkt aus der Artikulationstheorie ableiten. Diese zweite Definition zeigt eine erstaunliche Analogie zur Zuverlässigkeit von Parallelsystemen, so dass dem abstrakten Begriff der Robustheit eine konkretere Vorstellung zugeordnet werden kann. Darüber hinaus wird auf der Basis der Aurora-2 Datenbasis ein Experiment vorgeschlagen, mit dem das dazugehörige Maß bereits auf der Merkmalsebene für die unterschiedlichen *VAMIG*- Ausprägungen ermittelt werden kann. Zudem erfolgt auf der Basis der hier gewonnenen Einsichten eine Selektion der aussichtsreicheren *VVM*.

Kapitel 9

Um zu untersuchen, ob die Robustheit eines auditives Modells über einen weiten Störbereich erhalten bleibt, kann man das Training der *HMM* ausschließlich auf ungestörten Daten durchführen und anschließend bei unterschiedlichen Störszenarien und Störintensitäten testen. Diese Erkennexperimente für Ziffernketten für ausgewählte *VVM* finden wieder unter Verwendung der Aurora-2 Datenbasis statt.

Kapitel 10

In diesem Kapitel werden zunächst die Schlussfolgerungen bezüglich der erhobenen Hypothesen zur Bindungseigenschaft und der Einbeziehung von Vorhersagen zur Auflösung von Mehrdeutigkeiten zusammengefasst. Neben der experimentellen Bestätigung der beiden Hypothesen für die *PAS-CAS*-Modelle, führte auch die Berücksichtigung des Redundanz- und Kontinuitätsprinzip bei den hochauflösenden *PAS-CAS*-Modellen zu einer deutlichen Verbesserungen der Erkenngenauigkeit.

Der nachfolgende Ausblick hinsichtlich der Einbindung von weiteren strukturellen Informationen in den höheren Ebenen der Sprachverarbeitung basiert auf jüngeren Arbeiten, in denen die hierarchische Organisation des menschlichen Gedächtnissystems und die darauf optimierte bidirektionale Informationsverarbeitung berücksichtigt wird. Dabei kommt dem Vorhersageprinzip eine elementare Bedeutung zu. Darüber hinaus können die hier auftretenden Prozesse Codierung/Decodierung und Erkennung/Synthese bei hierarchischen Systemen als aufeinander aufsetzende Prozesse interpretiert werden, die einheitlich beschrieben werden können. Die grundlegende Idee eines hierarchischen Systems besteht aber in der Abstraktion. D.h. dass in höher liegenden Schichten weiter auseinanderliegende Korrelationen oder Abhängigkeiten erfasst werden können, mit denen die darunter liegenden Schichten mit näher beieinander liegenden Abhängigkeiten gesteuert werden können. Diese zusätzliche strukturelle Information hat bisher nur einen begrenzten Zugang in die Methoden der Spracherkennung gefunden.

Mit einer Schlussbemerkung werden die gewonnenen Einsichten und Ergebnisse noch einmal zusammengefasst und die Arbeit damit abgeschlossen.

2 Motivation

Der Artikulationstheorie ist eine über mehrere Jahrzehnte andauernde Experimentierphase in einem der berühmtesten Labore der Welt (S. Shannon war etwa zur selben Zeit in den Bell-Labs tätig) vorausgegangen. Frühzeitig erkannte H. Fletcher sowohl die Bedeutung von Kontext in einer Sprache als auch die Notwendigkeit störende Einflüsse in die Modellbildung einzubeziehen.

Die wichtigsten Folgerungen für ein Modell zur Erkennung von Sprache sind folgende: Offenbar verarbeitet das *HSR* voneinander unabhängige Teilbandsignale, deren Merkmale aber erst auf einer höheren Ebene für den nachfolgenden Erkennprozess zusammengefügt werden. Dies schließt prinzipiell die Möglichkeit ein, dass die phonetische Information in den einzelnen Teilbändern auch asynchron vorliegen kann. Eine weitere Konsequenz sowohl für das Modell, als auch für die Erstellung der Testdatenbasis ist der Einfluss von Kontext in der Sprache. Nach der Modellvorstellung von H. Fletcher kann die Sprachproduktion als eine Quelle verstanden werden, welche lautsprachliche Symbole erzeugt, die über einen gewissen Zeitraum voneinander abhängig sind. Eine derartige Quelle besitzt eine geringere Entropie als eine Quelle, deren Symbole unabhängig voneinander erzeugt werden. Damit die grundlegenden Mechanismen des menschlichen Erkennsystems unabhängig vom Einfluss des Kontext analysiert werden können, hat H. Fletcher in seinem Testkorpus ausschließlich sinnlose Silben, d.h. sinnfreie Kombinationen aus Vokalen und Konsonanten verwendet. Somit liegt prinzipiell eine Quelle mit maximaler Entropie vor. Für die Modellbildung folgt daraus, dass in den verschiedenen Ebenen des Modells (Phonem, Wort bzw. Satz), Information über den Kontext hinzugefügt wird, und somit schrittweise eine Reduktion der Entropie stattfindet.

In den nachfolgenden Jahren, in denen immer wieder neue Sprecher und Störsituationen einbezogen wurden, konnte eine bemerkenswerte Übereinstimmung zwischen Modell und Wirklichkeit bei unterschiedlichsten Geräusch- und Kanalbedingungen festgestellt werden. Darüber hinaus bietet die Theorie aber vor allem ein geschlossenes Modell mit dem die Artikulation – der Prozentsatz richtig erkannter Lauteinheiten - von Phonemen, Silben, Worten und Sätzen über einen weiten Bereich verschiedener Störungen vorhergesagt werden kann. Eine elementare Voraussetzung für die Vorhersage ist die Kenntnis des SNR in den einzelnen Artikulationsbändern. Eine unmittelbare Folge dieser Kenntnis ist zunächst einmal die Bestimmung der Artikulation von Phonemen. Davon ausgehend kann die Artikulation auf höheren Ebenen bestimmt werden. Bereits hier zeigt sich, welch große Bedeutung eine robuste Phonemerkennung für die menschliche Spracherkennung hat. Vergleicht man die heutigen Phonemerkennraten mit der menschlichen Performanz, so zeigen sich deutliche Nachteile gegenüber dem biologischen Vorbild. Dies gilt insbesondere dann, wenn die Erkennung in einem mit akustischen Störungen behafteten Umfeld stattfindet. Das lässt darauf schließen, dass wichtige Mechanismen in den aktuellen Modellen noch nicht berücksichtigt sind. In *[Allen-94]* wurden einige Ansätze vorgestellt, mit denen diese Probleme angegangen werden können und denen es lohnt nachzugehen. Ziel und zugleich Motivation dieser Arbeit ist es daher, relevante auditive Wirkprinzipien zu identifizieren und diese in einem möglichst recheneffizienten d.h. praxistauglichen Modell zu vereinen. Die Auswahl der auditiven Wirkprinzipien richtet sich dabei nach den wesentlichen Aspekten der Artikulationstheorie.

2.1 Die Artikulationstheorie

Eine wesentliche Motivation der vorliegenden Arbeit stellt die Artikulationstheorie dar, welche in Jahren 1918-1950 an den Bell Labs von einem Team unter der Leitung von H. Fletcher entwickelt wurde. In *[Allen-94]* wurden die Ideen und Resultate dieser Theorie unter Berücksichtigung bereits etablierter Theorien in einem neuen Gewand vorgestellt. Eine Zusammenfassung der wichtigsten Aussagen der Artikulationstheorie und weitere Einzelheiten findet man im Anhang B.

Die Artikulationstheorie wurde vor dem Hintergrund entwickelt, die Sprachverständlichkeit in unterschiedlichen Hörsituationen unabhängig vom Inhalt der übertragenen Nachricht und unabhängig von den individuellen Fähigkeiten von Sprecher- Hörer Paaren beurteilen zu können. Konkrete Hörsituationen werden neben dem Pegel des Sprachsignals vor allem durch den Pegel von Hintergrundstörungen und den durch den Kanal verursachten Dämpfungen und Verzerrungen des Signals charakterisiert, sie bestimmen die erreichbaren Erkennraten. Die Artikulationstheorie beschreibt die Sprachverständlichkeit bei unterschiedlichen Hörsituationen mit einer Variable, welche als Artikulationsindex bezeichnet wird. Die Beziehung zwischen einer konkreten Hörsituationen und dem zugehörigen Artikulationsindex ist durch das Artikulationsindex-Modell gegeben.

Die Information, die durch das Sprachsignal übertragen wird, ist auf bestimmte Sprachfrequenzen bzw. Teilbänder verteilt. Die der Artikulationstheorie zugrunde liegende Idee besteht nun darin, eine Transformation anzugeben, mit der die Information in den Teilbändern additiv zusammengefasst werden kann.

Dieser Ansatz fordert implizit, dass die Informationen in den Teilbändern unabhängig voneinander ihren Beitrag zur Sprachverständlichkeit leisten. In *[Fletcher-53]* wurde diese Forderung für ein Multiband-Modell wie folgt formuliert: Die Phonemfehlerrate ist gleich dem Produkt der Teilfehlerraten in den einzelnen Sprachbändern.

$$e = \prod_i e_i \tag{2.1.1}$$

Das bedeutet, dass in einem Multiband- Modell der Gesamtfehler immer kleiner ist als der kleinste Teilbandfehler. Die Teilbandfehler werden durch das SNR in den Teilbändern bestimmt $e_i = e_{\min}^{SNR_i}$ (siehe Anhang B). Formuliert man (2.1.1) mit der Phonemerkennrate $s = 1 - e$ und logarithmiert diesen Ausdruck, erhält man die gewünschte Additivität:

$$\log_{10}(1-s) = \sum_i \log_{10}(1-s_i) \tag{2.2.2}$$

Der zwischen Artikulationsindex und Hörsituation bestehende Zusammenhang wird auch als Artikulationsmodell bezeichnet:

$$AI = -k \log_{10}(1-s) = -k \sum_i A_i \tag{2.2.3}$$

Die Konstante *k* wird dabei so eingestellt, dass der Artikulationsindex für $SNR \geq 30dB$ den Wert eins annimmt.

Die von H. Fletcher verwendete Methode ähnelt einem psychophysikalischen Experiment, welches unter dem Namen ***D**iagnostic **R**hyme **T**est* bekannt ist. Mit diesem Sprecher-Hörer Experiment kann untersucht werden, wie gut ein Hörer phonetische Information wahrnehmen kann. Zunächst erkannte H. Fletcher und sein Team, dass die Testpersonen keinen Gebrauch von Kontextwissen bzw. höherem Wissen machen sollten. Daher wurde eine Datenbasis erzeugt, welche ausschließlich aus sinnlosen Silben bestand. Die Struktur dieser Silben ist durch die Abfolge Konsonant-Vokal-Konsonant festgelegt. Nachdem der Sprecher eine der sinnlosen Silben gesprochen hatte, sollte der Hörer aus einer Liste auswählen was er gehört hatte. Der Prozentsatz der korrekt erkannten sinnlosen Silben wurde Artikulation genannt.

Dieses Experiment wurde über einen außergewöhnlich langen Zeitraum für sich immer wieder ändernde Hörsituationen durchgeführt. Dabei wurde in den ersten Jahren lediglich das SNR modifiziert, später wurde zusätzlich eine Tiefpass-Hochpassweiche eingesetzt, um auch den Einfluss der Teilbänder auf die Sprachverständlichkeit untersuchen zu können.

Die wesentlichen Ergebnisse des Experiments können wie folgt zusammengefasst werden:

- Die drei Bestandteile (Phone) der sinnlosen Silbe werden unabhängig voneinander analysiert, d.h. das Phon scheint die elementare Einheit von Sprache zu sein.
- Phone (d.h. die akustische Realisierung eines Phonems) werden in unabhängigen Artikulationsbändern (Frequenzbändern) verarbeitet und deren lokale Merkmale bestimmt.
- Ein Phone wird durch die Beziehungen der lokalen Merkmale zueinander bestimmt. Über die Art der Beziehungen der lokalen Merkmale untereinander ist zunächst noch nichts bekannt.
- Die Artikulation hängt vom SNR in den Teilbändern ab, nicht jedoch von den Teilbandenergien.

Auf der Basis seiner langjährigen Experimente gelang es H. Fletcher ein Modell zu entwerfen, mit dem es möglich war, ausgehend von den Teilband-SNR die Phonemartikulation, die Silbenartikulation und die Wortartikulation vorherzusagen. Dieses Modell wurde für eine Vielzahl von Kombinationen unterschiedlicher Kanalparameter gestestet und zeigte eine beeindruckende Genauigkeit über einen weiten Bereich unterschiedlicher Hörsituationen *[Fletcher-53]*. Obwohl die ursprüngliche Motivation von H. Fletcher darin bestand, die Sprachqualität von Telefonkanälen beurteilen zu können, hat die Artikulationstheorie doch einige Aspekte aufzuweisen, die auch für das Gebiet der automatischen Spracherkennung relevant sind.

Mit der Veröffentlichung des Artikels von *[Allen-94]* nahm das Interesse an dieser Theorie wieder deutlich zu. Vor allem Multiband-Ansätze wurden stärker untersucht. Auch die Fragestellung wie Teilbandmerkmale zu einem Phon zusammengesetzt werden müssen, führte zu einer Reihe weiterer Untersuchungen. So wurden die Korrelationen zwischen den Teilbändern *[Bilmes-98]* als auch die zeitlichen Übergänge von Phonemen in den einzelnen Teilbändern untersucht *[Mirgafori-99], [Thomlinson-97], [Nilsson-00]*.

Die Bedeutung der Phonemerkennung wird bereits aus dem ersten der oben genannten Punkte deutlich. Selbst moderne ASR-Systeme sind derzeit noch weit von akzeptablen Phonemerkennraten entfernt. Mit dem nächsten Kapitel wird daher das Ziel verfolgt, auditive Wirkprinzipien zu identifizieren, die ihre Entsprechung in den oben genannten Punkten finden. Wenn die Phoneartikulation im Wesentlichen durch das SNR in den Teilbändern bestimmt wird, so müssen Mechanismen gesucht werden, mit denen eine Verbesserung der Teilband- SNR möglich ist. Eine weitere Forderung besteht darin Mechanismen zu finden, mit denen die Beziehungen der lokalen Merkmale untereinander modelliert werden können.

2.2 Auditive Modelle in der Spracherkennung

In den letzten Jahren sind fast alle klassischen Signalanalysetechniken wie *STFT*, Cepstrum oder die *LPC*- Technik um Aspekte der psychoakustischen Wahrnehmung erweitert worden (*PLP*, *MFCC*, oder *RASTA*). All diese Methoden basieren auf der Technik der Kurzzeit-Fourieranalyse, bei der das Signal in sich überlappende gleich große Blöcke unterteilt wird.

Dynamische Aspekte wurden meist durch die Differenzen zwischen aufeinander folgenden Blöcken durch Geschwindigkeits- oder Beschleunigungsparameter (Δ*MFCC*, $\Delta\Delta$*MFCC*) berücksichtigt. Nachdem das Verhalten der Basilarmembran, der inneren Haarzellen und den Nervenfasern genauer studiert worden konnte, entstanden zunächst einige einfache auditive Modelle mit denen die Transformationen denen das akustische Signal unterworfen wird, nachgebildet werden konnten. Obwohl diese Transformationen die physikalische Realität nur approximieren konnten, waren sie bereits geeignet um diejenigen Aspekte des Sprachsignals zu extrahieren, welche für die Analyse und Erkennung von Sprache relevant sind *[Wang-94a], [Wang-94b]*.

Fortgeschrittene Auditive Modelle berücksichtigen nicht nur auf der Wahrnehmung basierende Aspekte sondern auch Aspekte der Sprachproduktion sowie nichtlineare und dynamische Sprachcharakteristika *[Seneff-84], [Wang-93]*. Insbesondere wenn das Sprachsignal von Störgeräuschen überlagert ist, zeigt sich die Überlegenheit solcher Modelle gegenüber den klassischen Signalverarbeitungsmethoden. Eine Eigenschaft des von *[Wang-94a]* vorgeschlagenen Modells ist die Selbstnormalisierung des auditiven Spektrums. Das bedeutet, dass ein relatives Spektrum berechnet wird, welches eine gewisse Invarianz gegenüber additiven Störungen aufweist.

Bei weißen flachen Störungen können robuste Spektren berechnet werden, ohne dass eine spezielle Rauschunterdrückungsstufe verwendet werden muss. Eine biologisch orientierte Methode zur Berechnung von selbstnormalisierten Spektren stellen *LIN* dar. Neben der Selbstnormalisierung dient dieses Netzwerk auch der Schärfung von spektralen Spitzen bzw. der Extraktion von dominanten Periodizitäten. Ein wesentlicher Bestandteil des auditiven Modells von *[Seneff-84]* ist das Haarzellen-Synapsen-Modell. Dieses nichtlineare Modell ist allerdings ausgesprochen rauschempfindlich was zur Folge hat, dass der zur Verfügung stehende dynamische Bereich erheblich einschränkt wird. Da periodische Strukturen trotzdem erhalten bleiben, wurden die Werte der Teilband- AKF an den Stellen $\tau_i = CF_i^{-1}$ berechnet, die Größe dieser Werte gibt Auskunft über Periodizitäten in diesem Band. Mit dieser Methode konnte eine robuste spektrale Darstellung insbesondere von vokalischen Spektren erreicht werden.

Eine effiziente Implementierung zur Berechnung der Teilband-AKF wurde von *[Kajita-95]* vorgeschlagen. Eine Erweiterung des Ansatzes *[Kajita-98]* sieht eine Summe von exponentiell gewichteten Autokorrelationswerten bei Vielfachen von τ_i vor. Die mathematische Analyse dieses Ansatz macht deutlich, dass die AKF- Methode und die *LIN*- Methode äquivalent sind *[Kajita-98]*. Beide Methoden dienen der Schärfung der spektralen Spitzen des auditiven Spektrums und damit der Verbesserung des lokalen SNR.

Auch die Robustheit des *EIH* Spektrums *[Githza-94]* kann mit diesem Mechanismus erklärt werden. An dieser Stelle soll nicht unerwähnt bleiben, dass die Untersuchungen zur Erkenngenauigkeit des *EIH* in ähnlicher Weise wie die Experimente von H. Fletcher durchgeführt worden sind. In diesem Fall wurde ein *DRT* sowohl mit Testpersonen als auch mit einem automatischen Erkennsystem auf der Basis des *EIH*, der Fourieranalyse und der *LPC*- Technik durchgeführt. Bei derartigen Tests wird Kontextwissen bzw. höheres Wissen weitgehend ausgeblendet, so dass ein fairer Vergleich der verschiedenen Analysemethoden mit der menschlichen Erkennleistung möglich wird. Mit dieser Testphilosophie kann man zum einen unterschiedliche Methoden der Merkmalextraktion miteinander vergleichen, zum anderen können diese Methoden aber auch direkt mit der menschlichen Erkennleistung verglichen werden.

Die verwendete *DRT*-Datenbasis bestand aus 96 verwechselbaren Wortpaaren vom Typ *Konsonant-Vokal-Konsonant*, die sich in nur in ihren Anfangskonsonanten unterscheiden. Die Wortpaare wurden in sechs unterschiedliche phonetische Kategorien eingeordnet, wobei die Anfangskonsonanten so aufgeteilt wurden, dass pro phonetischer Kategorie 16 Wortpaare zur Verfügung standen. Der *DRT* liefert als Ergebnis eine Verwechslungsmatrix der Anfangskonsonanten bzw. eine Verwechslungsmatrix von phonetischen Kategorien. Bei diesen Experimenten wurde deutlich, wie empfindlich die klassischen Signalverarbeitungsmethoden auf Rauschstörungen reagieren.

[Perdigao-97] zeigte durch Experimente, dass auch das Haarzellen-Synapsen-Modell unter bestimmten Bedingungen zur Robustheit beitragen kann. Insbesondere bei schnellen Änderungen des Signals führt die Adaptionseigenschaft zu einer Schärfung der Signalübergänge und somit ebenfalls zu einer Verbesserung des lokalen SNR.

Außerdem wurde ein Verfahren zur Rauschunterdrückung vorgestellt, bei dem der Signalpegel von sprachfreien Segmenten unabhängig vom Kanalindex und unabhängig vom Rauschpegel konstant auf dem Niveau einer festen Maskierungsschwelle gehalten werden kann *[Perdigao-99]*. Ein derartiges Verfahren sorgt dafür, dass nach der Rauschunterdrückung ein flaches vom Rauschpegel unabhängiges Störspektrum vorliegt. Somit wird auch hier ein relatives auditives Spektrum - ein selbstnormalisiertes Spektrum - berechnet. Ein Nachteil des Perdigao-Verfahrens ist die begrenzte Auflösung, welche die Anwendung von *STFT*-Analyseverfahren mit sich bringt. Dieses Problem kann durch mehrfachauflösende Analyseverfahren überwunden werden. Bei dieser Methode steigt die zeitliche Auflösung mit zunehmender Mittenfrequenz des Analysefilters. Über viel versprechende Erfahrungen auf dem Gebiet der Spracherkennung konnten zuerst *[Jabloun-95]* und später auch *[Kryzse-99]* berichten. Die Autoren verwendeten eine Signalzerlegung basierend auf Wavelet-Paketen, welche als biorthogonale Filterbank mit Perfekter Rekonstruktion implementiert werden kann.

2.3 Weiterführende Modellierungen

In jüngster Zeit wurden einige auditive Modelle publiziert, die neben einer Analysefilterbank auch Modulationsfilterbänke für jedes einzelne Teilband vorsehen *[Dau-99], [Shamma-03], [Elhilali-06]*. Für die echtzeitfähige Spracherkennung sind solche Modelle allerdings noch nicht geeignet. Die Bindungseigenschaft wurde nach derzeitigem Wissensstand auch nur bei *[Elhilali-06]* implementiert. Ähnliche Einschränkungen gelten für die Ansätze der *Auditory Scene Analysis* Theorie. Die folgende Abbildung zeigt, wie zukünftige biologisch motivierte *VVM* aussehen könnten und wo die darin auszunutzenden Informationen im Frequenzbereich angesiedelt sind.

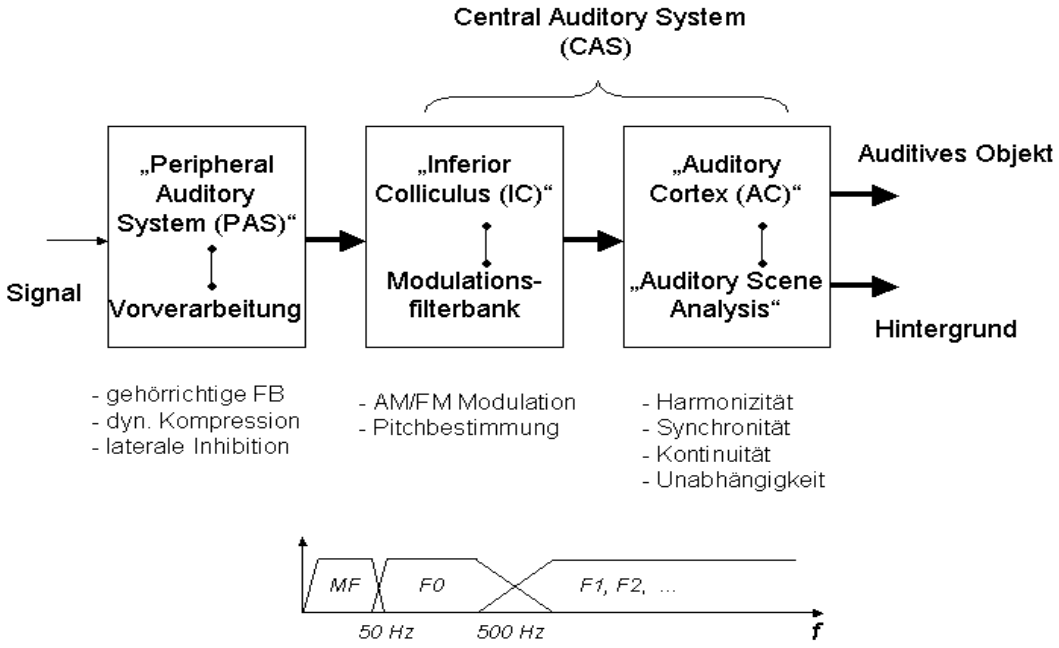

Abbildung 2.1 ***Biologisch motivierte Merkmalsextraktion. Im Bereich bis etwa 50 Hz findet man die AM/FM – Komponenten, die mit Modulationsfilterbänken gefunden werden können. Der Bereich bis 500 Hz dient der Pitchbestimmung, im darüber hinaus gehenden Frequenzbereich findet man die Formanten.***

Wenn sich diese Art von Merkmalsextraktion für die Automatische Spracherkennung realisieren lässt, so könnte dies zu kompakten Merkmalen mit einem höheren Informationsgehalt führen. Weitere Fortschritte, sind angefangen von der Geräuschreduktion über die robuste Parameterschätzung (Formanten, Sprachgrundfrequenz, Stimmhaftigkeit) bis hin zur Bildung von Auditiven Objekten zu erwarten *[Elhilali-06].*

3 Das Auditorische System

Durch immer leistungsfähigere, interdisziplinär angewandte Untersuchungsmethoden der Biophysik, Psychophysik und nicht zuletzt der Neurowissenschaften hat sich in den letzten Jahren das Bild von der auditiven Informationsverarbeitung gewandelt. Wurde in den Anfängen der physiologischen Akustik und Psychoakustik noch die Vorstellung von einer passiven Informationsaufnahme durch den Hörsinn verbreitetet, so wird heute der Prozess des Hörens als ein strukturierender, Bedeutung generierender Vorgang angesehen, bei dem Ohr und Gehirn zusammen aktiv die Hörempfindung hervorbringen. Die Beschreibung des Hörsystems und der Mechanismen der Verarbeitung von Schallereignissen erfolgt in der Literatur aus pragmatischen Gründen häufig anhand von aufsteigenden Kerngebieten der Hörbahn. Dabei sollte aber beachtet werden, dass zwischen nahezu allen Kerngebieten der Hörbahn auch absteigende Verbindungen existieren, über deren Bedeutung allerdings wenig bekannt ist. Die Ausführungen über die peripheren Anteile der Hörbahn nehmen einen relativ breiten Raum ein. Dies spiegelt sowohl den Kenntnisstand auf diesem Gebiet wieder und erklärt andererseits auch die Existenz von technisch realisierbaren Modellen. Anschließend werden die in der Literatur zur Sprachverarbeitung wohl bekanntesten Modelle detailliert vorgestellt, dabei wird nun der Schwerpunkt auf die Herausarbeitung der wichtigsten derzeit technisch umsetzbaren Wirkprinzipien gelegt. Dem schließt sich die Darstellung der zentralen Anteile der Hörbahn an, hier wird allerdings lediglich der Auditorische Cortex und der Colliculus Inferior beschrieben.

3.1 Das Periphere Auditorische System (PAS)

In diesem Abschnitt wird nachfolgend ein kurzer Abriss über den Aufbau des Gehörs, den Informationsfluss und die Signalverarbeitung im Innenohr gegeben.

Aufbau

Das Gehör besteht aus Außen- Mittel- und Innenohr. Das Außenohr hat eine Schutzfunktion, dient aber auch der Verstärkung und Lokalisation von Schallquellen. Das Mittelohr sorgt für den Druckausgleich, dient ebenfalls als Verstärker aber auch zur Anpassung der unterschiedlichen Schallwiderstände an die unterschiedlichen Übertragungsmedien Lymphflüssigkeit, Knochen und Luft. Würde dies nicht geschehen, dann würde ein Großteil der einfallenden Schallwellen an den Stoßstellen reflektiert werden und somit Information verloren gehen. Die Umwandlung und Kodierung der Lautstärke- und Frequenzinformation in entsprechende neuronale Impulse erfolgt im Innenohr.

Informationsfluss

Die Schallwellen der Luft werden von der Ohrmuschel in den Gehörgang geleitet, an dessen Ende sie das Trommelfell in Schwingungen versetzen. Eine Brücke aus drei Knöcheln (Hammer, Amboss und Steigbügel) leitet die Schwingungen zum Eingang des Innenohrs, das mit Lymphflüssigkeit gefüllt ist, die sich dann im Takt des letzten der drei Knöchelchen bewegt. Die Bewegung der Flüssigkeit reizt je nach Frequenz ganz bestimmte Sinneszellen, die auf einer länglichen zu einer Schnecke aufgerollten Membran, der so genannten Basilarmembran, angeordnet sind. In der Hörschnecke (Cochlea) findet die Umwandlung der mechanischen Energie in elektrische Energie statt, diese wird zum Informationstransport über die Nervenfasern benötigt. Die Nervenfasern werden zum Hörnerv gebündelt, über den die Information schließlich zur Großhirnrinde (Cortex) gelangt.

Signalverarbeitung im Innenohr

Breite und Elastizität der Basilarmembran sind längs der Hörschnecke verschieden. Am dünneren Ende ist die Membran sehr breit und schlaff (geringste Elastizität), am dickeren Ende ist die Membran schmal und steif (stärkste Elastizität). Wie schnell sich eine Schallwelle ausbreitet, hängt zum einen von der Masseverteilung und zum anderen von der der Elastizität der Basilarmembran ab. Diese Eigenschaften der Basilarmembran sind dafür verantwortlich, dass eine Schallwelle mit konstanter Frequenz immer langsamer wird, bis die Geschwindigkeit an einem bestimmten Punkt auf Null gesunken ist. Ganz in der Nähe dieses Punktes schwingt die Basilarmembran am stärksten auf und ab, so dass die Welle ihre Energie verbraucht und gestoppt wird. Die Geschwindigkeit der Schallwelle ist aufgrund von Schalldispersion nicht konstant, sie hängt von der Frequenz der Schallquelle und der Dämpfung durch das Medium ab. Je kleiner die Frequenz ist, desto kleiner ist die Ausbreitungsgeschwindigkeit, desto geringer ist auch die Schalldämpfung. Hohe Töne überholen demnach im Innenohr die tiefen Töne, aufgrund der höheren Dämpfung kommen die hohen Töne jedoch nicht soweit wie die tiefen Töne.

Die Ausbreitung der aufgezwungenen Schwingungen erfolgt in Form einer Wanderwelle, die sich an einem bestimmten Punkt totlaufen. Dies führt zu einer lokalen Erregung der Basilarmembran. An der Stelle wo sie am stärksten schwingt, werden die auf der Basilarmembran befindlichen Haare ausgelenkt. Die Stärke der Auslenkung steuert den Ionenstrom in die Haarzelle. Dieser Ionenstrom erzeugt elektrische Potentiale entlang der Haarzellenmembran. Diese Potentiale breiten sich über die Nervenfasern zum zentralen auditorischen System aus.

3.2 Eigenschaften von PAS Modellen

Psychoakustische Wahrnehmung

Für viele Höreigenschaften ist nicht die Frequenz das Wesentliche, sondern die Zahl der Haarzellen, die bei der Detektion des akustischen Ereignisses beteiligt sind. Psychoakustisch sind deshalb andere Skalen als die der Frequenz aussagekräftiger. Als Beispiel sei hier die Bark Skala genannt, 1 Bark entspricht einem Abschnitt der Basilarmembran von etwa 1.3 mm, das entspricht etwa einem Bündel von 150 in einer Reihe liegender Haarzellen. Die Bark Skala teilt Frequenzen in Frequenzgruppen (*engl. Critical Bands*) ein, auf dieser Skala haben dann alle Frequenzgruppen die gleiche Breite. Frequenzgruppen sind Frequenzbänder bestimmter Breite, welche bei einer Vielzahl von psychoakustischen Messungen in Erscheinung treten, und zwar dadurch, dass die Messergebnisse bezüglich einer Bandbreite unterhalb der Frequenzgruppenbreite deutlich anders ausfallen, als bei Bandbreiten oberhalb einer Frequenzgruppenbreite. Weitere gehörbezogene Frequenzskalierungen sind die Mel-Skalierung, die Spinc-Skalierung oder die ERB-Skalierung (***E**quivalent **R**ectangular **B**andwidth*) *[Terhardt-98].*

Die Lautstärke ist mit der Schallleistung einer Schallwelle verknüpft. Von den Schallwellen, die sich i.d.R. nach allen Seiten ausbreitet, erreicht nur ein Teil unser Ohr. Ein Maß für diese Teilleistung ist die Leistungsdichte oder Intensität der Schallwelle (Leistung pro Flächeneinheit). Da unser Ohr nicht gleich empfindlich auf alle Frequenzen anspricht, müssen die Intensitäten der verschiedenen Frequenzkomponenten mit unterschiedlichen Gewichtungsfaktoren multipliziert werden, bevor zusammengesetzte Klänge durch eine Gesamtintensität bewertet werden können.

Umfangreiche psychoakustische Messungen haben gezeigt, dass die wahrgenommene Lautstärke eines Tons von der Intensität und der Frequenz abhängt. Wichtige Kurven sind die Hörschwelle, die Isophone und die Schmerzschwelle. Die Hörschwelle bestimmt den Schallpegel bei der eine Frequenz gerade noch hörbar ist.

Die Schmerzschwelle dagegen markiert den Pegel, bei dem man Schall körperlich zu spüren beginnt und gesundheitliche Schädigungen einsetzen. Dazwischen liegende Kurven deren Frequenzen mit der gleichen Lautstärke empfunden werden, bezeichnet man als Isophone.

Die Lautheit ist ein absolutes Vergleichsmaß für Lautstärken, d.h. mit der Angabe der Lautheit hat man ein Maß zur Verfügung, mit dessen Hilfe man sagen kann, um wieviel mal lauter ein Ton im Vergleich zu einem anderen Ton klingt. Die Lautheit von Einzeltönen kann über die 3. Wurzel aus deren Schallintensität berechnet werden. Um die Gesamtlautstärke von zusammengesetzten Klängen zu bestimmen, ist eine Unterscheidung zwischen Tönen, die innerhalb bzw. außerhalb der kritischen Bandbreite liegen notwendig. Liegen Frequenzen um mehr als die kritische Bandbreite auseinander, so kann deren Lautheit einfach addiert werden. Bei Tönen innerhalb einer kritischen Bandbreite müssen erst deren Einzelintensitäten addiert werden, aus der Gesamtintensität kann dann die Teillautheit für das kritische Band ermittelt werden. Die Berechnung der Gesamtlautheit erfolgt dann wieder über die Summation aller Teillautheiten.

Ein weiterer Aspekt psychoakustischer Wahrnehmungen sind Modulationseffekte. Es ist bekannt, dass unser Gehör am empfindlichsten auf kleine Änderungen im Schall reagiert, wenn diese Änderungen etwa 4...8 -mal in der Sekunde erfolgen *[Eska-97]*. Diese Frequenzen werden bereits nicht mehr als Tonhöhen wahrgenommen (erst ab 20 Hz), sie entsprechen den typischen Änderungsgeschwindigkeiten bei der Artikulation von Sprachlauten. Die durchschnittliche Wortsilbendauer beträgt bspw. etwa 0.1 - 0.2 Sekunden.

Statische Eigenschaften von auditiven Modellen

Gemäß den psychoakustischen Wahrnehmungen können auditiven Modelle mit einigen statischen Eigenschaften ausgestattet werden. Neben der Verwendung von parallelen Bandpaßfiltern, deren Mittenfrequenzen und Bandbreiten durch die kritischen Bänder bestimmt sind, wird vor allem die Schallintensität der kritischen Bänder in die Lautheit transformiert. Als Beispiele für die in dieser Stufe verwendeten nichtlinearen Transformationen können die Bark- Skale, die Mel-Skale und die Intenstitäts-Lautheits-Transformation im *PLP*-Modell genannt werden.

Dynamische Eigenschaften

Es ist bekannt, dass einige dynamische Eigenschaften des peripheren auditorischen Systems für die Analyse und Verarbeitung von Sprache relevant sind. In jüngerer Zeit sind Veröffentlichungen bekannt geworden, in denen über Versuche berichtet wurde, diese Eigenschaften zu identifizieren und auch zu modellieren. Insbesondere die dynamischen Aspekte des Systems Haarzelle, Synapse und Nervenfaser konnten genauer untersucht werden. Die Systemantwort auf identische akustische Stimuli, die Ratenantwort sieht jedes Mal etwa anders aus. Daher wird die mittlere Pulsrate über verschiedene Realisierungen bestimmt, d.h. die mittlere Anzahl von Aktionspotentialen pro Zeiteinheit wird als Balkendiagramm aufgetragen. bestimmt. Dieses sogenannte Peri-Stimulus-Time-Histogramm gibt die typische Antwort des Systems nach Präsentation eines bestimmten Reizes wieder.

Experimente haben gezeigt, dass die Ratenantwort des Systems Haarzelle, Synapse und Nervenfaser innerhalb der ersten 15 ms nach Einsatz eines dauerhaft anhaltenden akustischen Stimuli am stärksten ist. Typischerweise beobachtet man danach zunächst einem sehr schnellen Abfall, gefolgt von einem langsameren Abfall der Ratenantwort auf ein stabiles Niveau. Dieses Phänomen wird in der Fachliteratur als Adaption bezeichnet.

Die Zeitdauer innerhalb der ein langsamer Abfall beobachtet werden kann, bezeichnet man als Refraktärzeit. Dies ist eine Zeitspanne innerhalb der das System nicht auf neue Reize reagieren kann. Die Refraktärzeit bestimmt also das zeitliche Auflösungsvermögen von zwei aufeinander folgenden akustischen Reizen.

Eng damit verbunden ist die Eigenschaft der Vorwärtsmaskierung, dieses Phänomen tritt auf, wenn ein vorangegangenes intensives akustisches Signal ein nachfolgendes schwächeres Signal verdeckt. Neben diesen beiden Eigenschaften sollte noch das Phänomen der Phasenkopplung genannt werden. Dieses Phänomen kann insbesondere bei Erregungen im tieffrequenten Bereich beobachtet werden, bei hohen Frequenzen tritt das Phänomen nicht auf. Daher spricht man von einem Synchronitätsverlust.

In *[Seneff-84]* wurde ein 3 stufiges Modell vorgestellt, welches die oben genannten Eigenschaften mit genügender Genauigkeit nachbildet. Auf dieses Modell wird im Abschnitt 3.4 genauer eingegangen.

*Modulationsmodelle und der **T**eager **E**nergy **O**perator (TEO)*

Zu denjenigen Modellen, welche dynamische Eigenschaften nachbilden, können auch diejenigen Modelle gezählt werden, welche Modulationseffekte berücksichtigen. Hier ist insbesondere das *RASTA*- Modell hervorzuheben *[Hermansky-94]*. In diesem Modell werden nach der Berechnung der logarithmierten Energien in den einzelnen Teilbändern die Zeit-Trajektorien innerhalb der Teilbänder mit einem Bandpass gefiltert, d.h. es werden diejenigen Frequenzkomponenten der Einhüllenden unterdrückt, welche sich langsamer oder schneller als mit der typischen Modulationsfrequenz von 4.. 8 Hz ändern. Da das vorgeschlagene Bandpass- Filter an der Stelle *z = 1* bzw. *f = 0* eine Nullstelle besitzt, werden die konstanten Anteile (diese sind dem Kanal zu zuordnen) unterdrückt. Die aus dem *RASTA*- Verfahren abgeleiteten Varianten *lin-log-RASTA* und *RASTA-PLP* kombinieren bereits dynamische und statische Eigenschaften, so werden neben der Verwendung einer gehörrichtigen Filterbank und der Intensitäts- bzw. Lautheits- Transformation auch Modulations- und Maskierungseigenschaften im Zeit- und Frequenzbereich ausgenutzt.

Das Sprachsignal kann hinsichtlich seiner Natur als ein Mehrfachkomponenten-Signal verstanden werden. Die Formanten d.h. die Resonanzfrequenzen des Vokaltraktes sind energiereiche lokal dominierende Signalkomponenten deren Frequenzen, Einhüllenden und Bandbreiten sich mit der Zeit ändern. Mit dem Modulationsmodel können solche lokalen Eigenschaften der Signalstruktur durch eine Summe von AM- und FM-Signalen beschrieben werden. Ein einzelner Formant kann demnach wie folgt beschrieben werden:

$$r(t) = a(t)\cos(\omega_c t + \omega_m \int_0^t q(\tau)d\tau + \theta) \tag{3.2.1}$$

Das Argument der Kosinusfunktion wird im Allgemeinen als Phasenfunktion bezeichnet:

$$\phi(t) = \omega_c t + \omega_m \int_0^t q(\tau)d\tau + \theta \tag{3.2.2}$$

Die Einhüllende des Formanten wird mit *a(t)* bezeichnet, ω_c entspricht der Trägerfrequenz, ω_m legt die maximale Frequenzabweichung von der Trägerfrequenz fest und θ ist ein konstanter Phasen-Offset. Für die frequenzmodulierende Funktion *q(t)* gilt dann:

$$|q(t)| \le 1 \tag{3.2.3}$$

Die Momentanfrequenz $\omega_i(t)$ des Formanten erhält man dann durch Differentiation der Phasenfunktion:

$$\frac{d\phi(t)}{dt} = \omega_i(t) = \omega_c + \omega_m q(t) \tag{3.2.4}$$

Bezieht man in die Signalbeschreibung alle Formanten ein, so kann das Sprachsignal als Linearkombination von AM- und FM-Komponenten in der folgenden Form dargestellt werden:

$$s(t) = \sum_k a_k(t) \cos \phi_k(t) \tag{3.2.5}$$

In dieser Darstellung sind die voneinander unabhängig variierenden Komponenten $\omega_i(t)$ und $a(t)$ die bestimmenden Größen. Über die Verwendung der AM-FM-Komponenten zur Bildung robuster Merkmale für die Automatische Spracherkennung wird bspw. in *[Dimitriadis-05b]* berichtet.

Der Einführung des ***T**eager **E**nergy **O**perator* geht nun auf physikalische Betrachtungen eines linearen Oszillators *[Kaiser-90]* zurück. Die Momentanenergie setzt sich zu jedem Zeitpunkt aus kinetischer und potentieller Energie zusammen, wobei diese nicht nur proportional zum Quadrat der Amplitude sondern auch proportional zum Quadrat der Momentanfrequenz ist. Diese Betrachtung berücksichtigt also ebenfalls die beiden bestimmenden Komponenten des Modulationsmodells. Der kontinuierliche *TEO* genügt formal der folgenden Beziehung:

$$\Psi_c\,[r(t)] \;\sim\; a(t)^2 \omega_i(t)^2 \tag{3.2.6}$$

Durch Bildung des RMS- Wertes liegt eine Methode vor, mit der man dem Produkt aus der AM- Einhüllenden und der FM- Momentanfrequenz folgen kann.

$$\Psi_c^{\,RMS} = \sqrt{\frac{1}{T}\int \Psi_c\,[r(t)]} \tag{3.2.7}$$

Mit der zeitdiskreten Version des *TEO* steht nun ein ausgesprochen einfacher Algorithmus zur Verfügung, um Modulationsaspekte in der Struktur des Sprachsignals erfassen zu können.

$$\Psi_d\,[r(n)] = r^2(n) - r(n+1)r(n-1) \tag{3.2.8}$$

Dieser Algorithmus bzw. der Kurzzeit- RMS Wert des *TEO* kann bei echten Teilbandmodellen zur Erfassung von AM- FM- Merkmalen verwendet werden. Der *TEO* wirkt gemäß der Artikulationstheorie nur innerhalb eines Teilbandes entlang der Zeitachse. Darüber hinaus wurde in *[Maragos-93]* aber noch eine weitere wichtige Eigenschaft des *TEO* gefunden, mit der dieser Operator einer ganz wesentlichen Forderung der Artikulationstheorie genügt. In bestimmten Situationen kann der *TEO* das lokale SNR verbessern. Hierzu betrachte man das SNR- Verhalten des *TEO* bei Überlagerung eines Nutzsignals mit weißem Rauschen $y(n) = x(n) + w(n)$. Für den Erwartungswert des gestörten Signals erhält man zunächst:

$$E\{\Psi[x(n)+w(n)]\} = E\{\Psi[x(n)]\} + \delta_w^2 \quad ; \quad \delta_w^2 = E\{w^2(n)\} \tag{3.2.9}$$

Der Erwartungswert des *TEO* in (3.2.9) entspricht der mittleren Leistung des zeitdiskreten TEO,

$$P_\Psi = \frac{1}{N}\sum_{i=1}^{N} \Psi_d[x(n)] \tag{3.2.10}$$

den RMS des *TEO* erhält man mit der Quadratwurzel aus der mittleren Leistung:

$$\Psi_d^{RMS} = \sqrt{\frac{1}{N}\sum_{n=1}^{N}\Psi[x(n)]} \tag{3.2.10}$$

Das SNR am Eingang und am Ausgang des Operators ist durch (3.2.11) gegeben:

$$SNR_I = \frac{P_{xx}}{\delta_w^2}; \qquad SNR_O = \frac{P_\Psi}{\delta_w^2} \tag{3.2.11}$$

Die Autoren in *[Maragos-93]* haben noch eine weitere wichtige Beziehung angegeben:

$$SNR_O \leq 2\sin^2(\Omega_x) SNR_I \text{ , } mit\ \Omega_x = 2\pi f / f_S,\ \Omega_x \in \{0..\pi\} \tag{3.2.12}$$

wobei das Gleichheitszeichen für Nutzsignale der Form $x(n) = A\cos(\Omega_x n)$ gilt, dann aber folgt hieraus:

$$SNR_O = SNR_I\,|_{\Omega=\pi/4,\Omega=3\pi/4}$$

$$SNR_O = 2\,SNR_I\,|_{\Omega=\pi/2} \tag{3.2.13}$$

D.h. der *TEO* ist ein Operator, mit dem in bestimmten Situationen das lokale SNR erhöht werden kann. Dies gilt wohl insbesondere für die *WT*, denn hier wird bekanntlich jedes Band auf das Frequenzintervall [0…π] unterabgetastet. Die Mittenfrequenz eines jeden *WT*-Kanals findet man immer an der Frequenz π/2. Durchläuft also ein mit weißem Rauschen gestörtes kosinusförmiges Signal die Mittenfrequenz, so verdoppelt sich das SNR am Ausgang des Operators. Dies ist aber eine der gewünschten Eigenschaften, welche eine Signalverarbeitungskomponente gemäß der Artikulationstheorie aufweisen sollte.

3.3 PAS Modell nach S. Shamma

Bei biologisch orientierten sensorischen Vorverarbeitungsstufen wird das Eingangssignal häufig durch eine mehrfachauflösende Transformation dargestellt. Die aus dieser Transformation hervorgehenden Teilsignale erfahren anschließend durch eine nichtlineare Kennlinie eine Kompression. Aufgrund von Sättigungseigenschaften des sensorischen Kanals und infolge der Wirksamkeit einer neuronalen Erregungsschwelle wird der dynamische Bereich eines derart transformierten Teilsignals begrenzt.

Eine mathematische Formulierung der zugehörigen Signaltransformationen, mit denen die peripheren Verarbeitungsstufen des Gehörs beschrieben werden können, wurde von *[Yang-92]* präsentiert. Im Wesentlichen durchläuft das Signal demnach die folgenden Stufen: Analyse, Signalwandlung und Reduktionsstufe (vergl. Abb. 3.3.1).

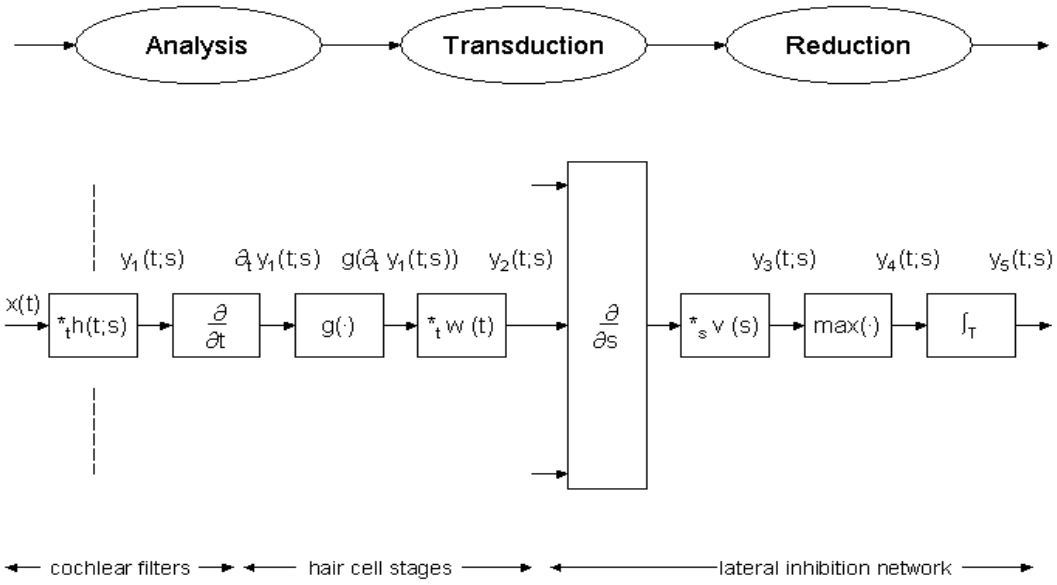

Abbildung 3.3.1 ***Mathematisches Modell der wesentlichen Verarbeitungsstufen im Peripheren Auditorischen System nach S. Shamma***

A) Diese Stufe berücksichtigt die Beziehung die zwischen den geometrischen Orten der Basilarmembran und dem Amplituden- und Frequenzinhalt des erregenden Signals besteht. D.h. die im Signal enthaltenen Frequenzen erregen die Haarzellen an verschiedenen Orten der Basilarmembran, dieser cochleare Mechanismus zur Trennung der eingehenden Signalfrequenzen wird als Bank von parallelen Bandpassfiltern modelliert. Die Mittenfrequenzen der auch als Critical- Band-Filter bezeichneten Bandpassfilter sind unterhalb von 800 Hz linear und dann ungefähr logarithmisch entlang der Frequenz verteilt. Die spektrale Intensität des akustischen Signals wird dabei als Summe der Intensitäten aller kritischen Teilbandfilter wahrgenommen. Die Intensitäten der kritischen Teilbänder werden über eine nichtlineare Transformation (meist wird der Logarithmus oder eine nicht ganzzahlige Potenz verwendet) in eine wahrgenommene die Lautheit überführt.

B) Die Wandlung der mechanischen Bewegung entlang der Basilarmembran in die Feueraktivitäten von erregten Haarzellen findet in der Signalwandlungsstufe statt. Dabei handelt es sich um einen dreistufigen Prozess: Der Ionenstrom durchläuft einen nichtlinearen Kanal um in die Haarzelle zu gelangen, ein Tiefpassfilter erzeugt ein langsam variierendes Signal in jedem kritischen Band.

C) Einen wichtigen Aspekt dieser Stufe stellt die Laterale Inhibition dar, im Wesentlichen wird mit diesem Mechanismus die spektrale Charakteristik des Signals geschärft.

Die Analyse der einzelnen Modellierungsstufen ermöglichte zum einen tiefere Einsichten hinsichtlich auditiver Wirkprinzipien zu erlangen, zum anderen gelang auch eine Erklärung bezüglich des häufig beobachteten robusten Verhalten gegenüber Rauscheinflüssen. Zunächst wird in der 1. Stufe die Ausbreitung von Wanderwellen entlang der Basilarmembran modelliert.

$$y_1(t;s) = h(t;s) *_t x(t) \tag{3.3.1}$$

Die nachfolgenden Transformationen der zweiten und dritten Stufe beschreiben die Umwandlung der mechanischen Energie der Wanderwelle in die neuronale Feueraktivität der Nervenfasern.

$$y_2(t;s) = g(\partial_t y_1(t;s)) *_t w(t) \tag{3.3.2}$$

Die nichtlineare Kompressionskennlinie ist in zwei lineare Operationen eingebettet. Vor der Nichtlinearität findet eine Ableitung der Teilbandsignale nach der Zeit statt, da die Stärke der Auslenkung der Haarzellen von der Geschwindigkeit der Wanderwelle an der entsprechenden Position abhängt. Der Nichtlinearität folgt ein Tiefpass, welcher den Abfluss der Ionen durch die Zellmembran modelliert.

$$g(u) = \frac{1}{1+e^{-\gamma u}} - \frac{1}{2} \tag{3.3.3}$$

An dieser Stelle sei erwähnt, dass die Nichtlinearität für $\gamma \to \infty$ in die Einheitssprung-Funktion übergeht.

$$g(u) = \begin{cases} 1 & ,u \geq 0 \\ 0 & ,u < 0 \end{cases} \tag{3.3.4}$$

In der 3. Stufe unterliegt das Signal einer weiteren Transformation, welche alle biologisch orientierten sensorischen Systeme gemeinsam haben. Diese Transformation wird als Laterale Inhibition bezeichnet und reduziert die korrelierten Feueraktivitäten zwischen den Teilsignalen. Die Dekorrelation führt daher zu einer Verstärkung von schnellen Änderungen entlang der Frequenzachse.

$$y_3(t;s) = \partial_s (g(\partial_t y_1(t;s))) *_t w(t) *_s v(s) \tag{3.3.5}$$

Schließlich erfolgt die auditive Darstellung des akustischen Signals durch eine Halbwellengleichrichtung (***H**alf **W**ave **R**ectification*) mit nachfolgender Glättung in Zeitrichtung. An dieser Stelle sei nochmals darauf hingewiesen, dass die Information welche in der Einhüllenden des Signals $y_3(t,s)$ kodiert ist, in der auditiven Darstellung erhalten bleibt.

$$y_4(t) = \max(y_3(t;s), 0) \tag{3.3.6}$$

$$y_5(t) = \frac{1}{T} \int_{t-T}^{t} y_4(\tau; s)\, d\tau \tag{3.3.7}$$

Im Folgenden wird das Modell hinsichtlich seiner Eigenschaften analysiert. Eine wesentliche Vereinfachung der mathematischen Analyse gelingt, wenn die Nichtlinearität im sogenannten *High-Gain* Grenzbereich betrieben wird, also die Einheitssprungfunktion verwendet wird. Berücksichtigt man in diesem Fall die Beziehung $\frac{dg(u)}{du} = \delta(u)$ kann das Signal am Ausgang der 2. Stufe wie folgt beschrieben werden:

$$y_3(t,s) = \left(\delta(\partial_t y_1(t,s))\partial_s \partial_t y_1(t,s)\right) *_t \; w(t) *_s \; v(s) \tag{3.3.8}$$

Nach *[Papoulis-02]* gilt:

$$\delta(f(t)) = \sum_{t_i \in Z} \delta(t_i)\partial_t f(t_i) \tag{3.3.9}$$

wobei mit

$$Z = \{t : \; f(t) = 0\} \tag{3.3.10}$$

die Positionen t_i der Nulldurchgänge der Funktion $f(t) = \partial_t(y_1(t,s)$ definiert sind. Diese Darstellung entspricht einem signalgetriebenen Abtastprozess, die Abtastzeitpunkte t_i werden durch die Nulldurchgänge der Funktion $\partial_t y_1(t,s)$ bzw. durch die Position der Extremwerte der Funktion $y_1(t,s)$ gegeben sind.

Die Funktion $y_4(t,s)$ berücksichtigt lediglich die positiven Werte, aus denen abschließend ein Kurzzeit- Mittelwert berechnet wird. Wenn die Nichtlinearität nicht vorhanden wäre, würde lediglich die Einhüllende von $y_3(t,s)$ übertragen werden. Die auditive Darstellung des Signals geht in diesem Fall in ein einfaches Kurzzeit-Spektrum über.

Der in Klammern gefasste Term in (3.3.8) kann dahingehend interpretiert werden, dass die Extremwerte der Funktion $y_1(t,s)$ mit den partiellen Ableitungen $\partial_s \partial_t y_1(t,s)$ - an den Stellen der Extremwerte - gewichtet werden.

$$S(t,s) = \frac{\partial_s \partial_t y_1(t,s)}{\partial_t^2 y_1(t)} \tag{3.3.11}$$

Der Ausdruck $\partial_s \partial_t y_1(t,s) = \left(\partial_t x(t)\right) *_t \left(\partial_s h(t,s)\right)$ stellt eine Wavelet-Transformation dar, hier wird allerdings die Ableitung des Eingangssignals einer Bank von Differentialfiltern $\partial_s h(t,s)$ zugeführt. Bei geeigneter Wahl der Cochlea-Filter sind die Differentialfilter sehr viel schmalbandiger als die der Cochlea-Filterbank. Da $\partial_t y_1(t,s)$ ein Bandpaßsignal ist, kann es durch Momentanamplitude und Momentanphase dargestellt werden:

$$\partial_t y_1(t,s) = a(t,s)\sin(\omega_{ds} + \phi_{ds}) \tag{3.3.12}$$

In *[Wang-94a]* wird gezeigt, dass das auditive Spektrum im Grunde durch das Verhältnis der Energien in den schmalbandigen Differentialfiltern und den Energien in den breitbandigen Cochleafiltern bestimmt wird, dies führt zu einem Selbstnormalisierungseffekt bei Störungen mit Gaußschen Rauschen. Dieser Effekt ist durch das Wirkprinzip der ***L**ateralen **I**nhibition* begründet und hat zur Folge, dass das auditive Spektrum unabhängig vom Rauschpegel in normalisierter Form vorliegt. Eine noch schärfere Formulierung der schmalbandigen Differentialfilter als Delta-Funktion an der Stelle der Mittenfrequenz ω_{cs} führt auf eine Darstellung, in der das auditive Spektrum im Wesentlichen durch die Beträge des Eingangsspektrums $|X(\omega_{cs})|$ bestimmt wird, dieses aber durch den Term *a(t,s)* normiert und mit dem Kosinus des Abstandes der dominanten Frequenz ω_{ds} zur Mittenfrequenz ω_{cs} gewichtet wird.

$$S(t,s) = \left[\frac{|\omega_{cs} X(\omega_{cs})|}{|\omega_{ds} a(t,s)|} \cos(\Delta\omega_s t + \Delta\phi_s) \right] \tag{3.3.13}$$

$$\Delta\omega_s = \omega_{cs} - \omega_{ds} \ , \quad \Delta\phi_s = \phi_{cs} - \phi_{ds} \tag{3.3.14}$$

Dies geht aus nachfolgender Betrachtung hervor: Mit $\Delta\omega_s = \Delta\phi_s = 0$ kann *a(t,s)* ungefähr als Konstante betrachtet werden, welche die Momentanenergie im zugehörigen Cochlea-Filter repräsentiert, die Einhüllende des auditiven Spektrums gibt demnach ein normalisiertes Maß der Stärke des Signals an der Stelle $|X(\omega_{cs})|$ wieder. Mit zunehmender Differenz von $\Delta\omega_s$ *bzw.* $\Delta\phi_s$ wird daher auch das Signal an der Stelle $|X(\omega_{cs})|$ zunehmend gedämpft. Dieser Aspekt des auditiven Spektrums wird später bei der Realisierung eines *LIN* zu berücksichtigen sein. Zunächst soll lediglich festgestellt werden, dass es für eine *LIN*- Realisierung notwendig ist, die dominante Frequenz im Teilband *s* schätzen zu können.

3.4 PAS Modell nach S. Seneff

Die Modellierung des peripheren auditorischen Systems durch das kombinierte *Synchrony-Mean-Rate*-Modell *[Seneff-84]* erfolgt in 3 Stufen (Abbildung 3.4.1). Die ersten zwei Stufen bilden die Transformationen des peripheren auditorischen Systems nach, mit der dritten Stufe werden für die Wahrnehmung relevante Informationen – wie die Lokalisierung phonetischer Übergänge oder dominante Periodizitäten zur Identifikation der Formanten – extrahiert. In der ersten Stufe wird das auf 8 kHz bandbegrenzte Sprachsignal einer barkskalierten Filterbank zugeführt. Die Bandbreite der Filter entspricht etwa 0.5 Bark. Anschließend durchlaufen die Bandpass-Signale das Haarzellen-Synapsen-Modell. Mit diesem nichtlinearen Modell wird die Umwandlung der mechanischen Bewegungsenergie in neuronale Feueraktivität simuliert. Der Ausgang dieser Stufe repräsentiert die Feuer-Wahrscheinlichkeit einer Gruppe von Nervenfasern als eine Funktion der Zeit.

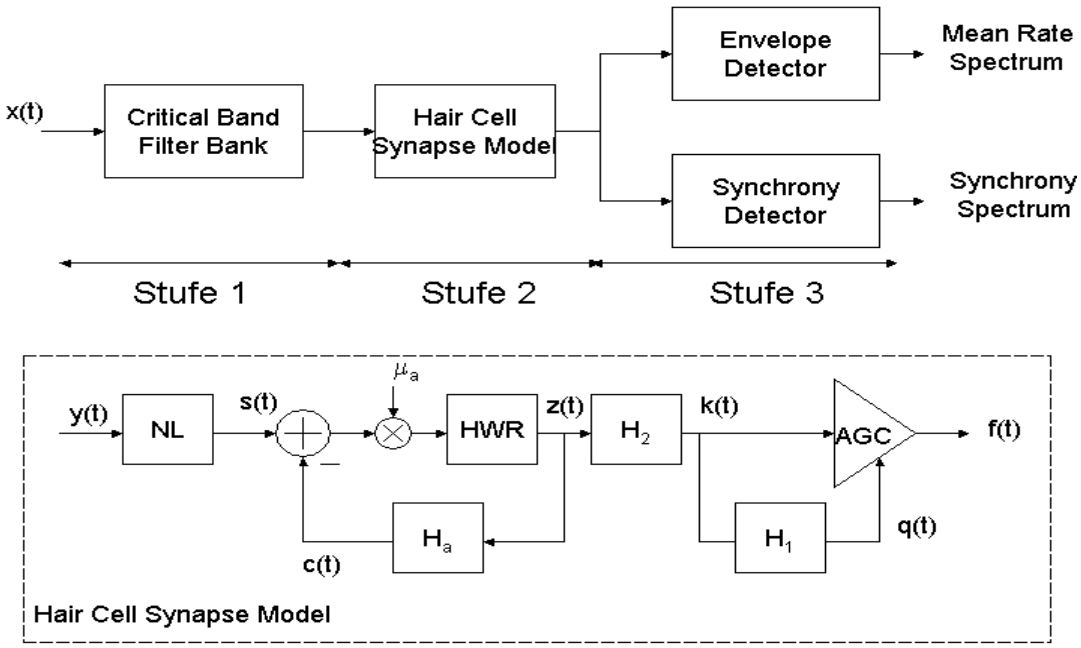

Abbildung 3.4.1 ***Das Modell von S. Seneff berücksichtigt vor allem die dynamischen Eigenschaften des Peripheren Auditorischen Systems, diese sind im Haarzellen Modell berücksichtigt.***

In der zweiten Stufe werden insgesamt vier verschiedene neuronale Mechanismen modelliert. Das direktionale Antwortverhalten der inneren Haarzellen wird durch die folgende nichtlineare Kompressionskennlinie simuliert.

$$s(t) = \begin{cases} 1 + A \tan^{-1}(B\, y(t)) & , y > 0 \\ e^{AB\, y(t)} & , y \le 0 \end{cases} \qquad (3.4.1)$$

Der nächste Block steuert das dynamische Antwortverhalten hinsichtlich plötzlicher Signaländerungen. Die so genannte Kurzzeit-Adaption berücksichtigt die dynamischen Eigenschaften eines Prozess, der für die Ausschüttung von Neurotransmittern in die synaptische Region zwischen Haarzelle und Nervenfaser verantwortlich ist.

Dieser eher langsame Prozess besteht aus zwei separaten Mechanismen welcher die Konzentration der Neurotransmitter beeinflusst und somit den nichtlinearen Kanal öffnet oder schließt.

$$\frac{dc(t)}{dt} - \mu_b c(t) = \begin{cases} \mu_a (s(t) - c(t)) & , (s(t) - c(t)) \geq 0 \\ 0 & , (s(t) - c(t)) < 0 \end{cases} \tag{3.4.2}$$

Der nächste Mechanismus berücksichtigt den Abfall der Synchronität des Antwortverhaltens von Nervenfasern wie er bei akustischen Stimuli mit hohen Frequenzen beobachtet wird. Dieser Effekt wird durch einen Tiefpass 4. Ordnung simuliert

$$H_1(z) = \left(\frac{1-\alpha}{1-\alpha\ z^{-1}} \right)^4 \quad , \alpha = \exp\left(-1 \Big/ f_s \tau_{LP} \right) \tag{3.4.3}$$

Eine schnelle AGC modelliert den sehr schnellen Abfall der Feueraktivität, welcher kurz nach dem Einsatz eines akustischen Stimuli beobachtet werden kann. Sie bildet den Abschluss der zweiten Stufe.

$$f(t) = \frac{k(t)}{1 + K\, q(t)} \quad , mit \;\; q(t) = \overline{k(t)} \tag{3.4.4}$$

Zusammen mit der Kurzzeit-Adaption simuliert die AGC den Refraktäreffekt. Dieser Effekt hat zur Folge, dass das Modell innerhalb der so genannten Refraktärzeit für neue Erregungen unempfindlich ist. Die durch die erste und zweite Stufe simulierten Eigenschaften des auditorischen Systems sollen an dieser Stelle nochmals genannt werden:

- Bark- Skalen Filterung,
- nichtlineare Kompressionskennlinie,
- Vorwärtsmaskierung,
- Refraktäreffekt,
- Verlust an Synchronität bei hohen Frequenzen

Die dritte Stufe enthält zwei unabhängig voneinander arbeitende Signalverarbeitungsblöcke. Im ersten Block - dem Einhüllenden Detektor - wird der Ausgang der 2. Stufe geglättet und abwärts getastet, dieser Block folgt der Einhüllenden der neuronalen Feuerrate und ist für die Lokalisierung phonetischer Übergänge geeignet. Aufgrund des Sättigungsverhaltens der nichtlinearen Kennlinien und dem damit verbundenen limitierten dynamischen Bereichs kommt es bei vokalischen Segmenten zu einer Verschmierung der Formantfrequenzen. Das liegt daran, dass diese Abschnitte ein höheres Energielevel besitzen und somit benachbarte Teilbänder in die Sättigung treiben, die Frequenzauflösung verschlechtert sich damit.

Der zweite Verarbeitungsblock wird als Synchronitätsdetektor bezeichnet. Trotz der Sättigungseffekte bei vokalischen Segmenten behalten die Teilbandsignale ihre periodische Struktur. Diese Struktur wird ausgenutzt und dominante Periodizitäten insbesondere die der Formanten werden hervorgehoben, Periodizitäten der glottalen Erregung dagegen werden reduziert. Aus dem Ausgangssignal der 2. Stufe eines jeden Teilbands wird zunächst ein zweites Signal mit der Verzögerungszeit $\tau_i = CF_i^{-1}$ erzeugt. Aus der Summe und der Differenz der beiden Teilsignale wird anschließend je ein Erwartungswert berechnet.

Ein weich begrenztes Verhältnis dieser beiden Erwartungswerte bildet den Ausgang des *i*-ten Synchronitätsdetektors.

$$GSD_i = \tan^{-1}\left(\frac{E\{|\ x_i(n) + x_i(n-\tau_i)\ |\} - \delta}{E\{|\ x_i(n) - x_i(n-\tau_i)\ |\}}\right) \qquad (3.4.5)$$

Mit diesem Block wird die Frequenzauflösung verbessert und eine Schärfung des auditiven Spektrums erreicht.

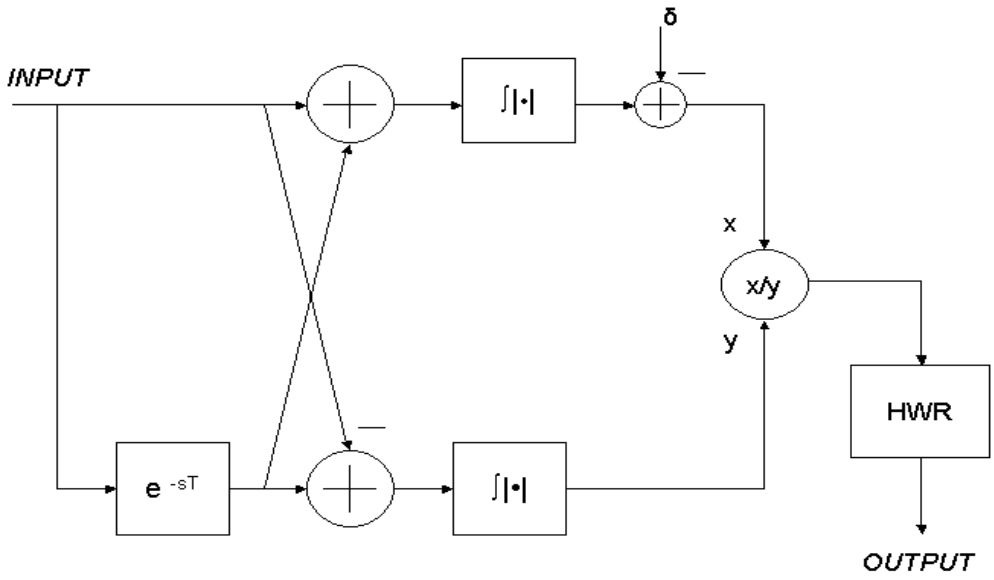

Abbildung 3.4.2: ***Der Generalized Synchrony Detector (GSD) dient der Verstärkung bzw. Schärfung dominanter periodischer Anteile. Diese Anteile bleiben trotz der Sättigungseigenschaften der nichtlinearen Kennlinie erhalten.***

Bei Experimenten mit realen Sprachsignalen konnte gezeigt werden, dass wesentliche Eigenschaften des Sprachsignals wie Pitch- und Formantfrequenzen nach Verarbeitung durch das Modell gut erhalten bleiben. Weiterhin konnten deutliche Kontrastunterschiede beim Übergang von einem phonetischen Segment zum nächsten beobachtet werden. Problematisch scheint der begrenzte dynamische Bereich des Modells zu sein, dieser wird vor allem durch die nichtlineare Kompressionskennlinie des Haarzellenmodells verursacht. Infolge dessen kommt es zu einer Verschmierung der Formantfrequenzen.

Die wesentlichen Charakteristika der dritten Stufe sind durch die folgende Aufzählung gegeben:

- Lokalisierung phonetischer Übergänge
- Extraktion dominanter Periodizitäten

Die Synchronitätsdetektion der 3. Stufe hat sich bei Experimenten mit weißen Rauschstörungen gegenüber der Einhüllenden- Darstellung als die robustere Darstellung erwiesen.

In *[Ali-00]* wird ein Ansatz vorgestellt, mit dem die Detektion der Formanten nochmals wesentlich verbessert wird und die durch die Berechnungsvorschrift nach S. Seneff auftretenden Artefakte wie z.B. zufällige spektrale Spitzen deutlich reduziert werden. Dem *i*-ten GSD-Element werden neben dem Signal des *i*-ten Filterausgangs auch die Signale benachbarter Filterausgänge präsentiert. Die Ausgänge des *i*-ten GSD- Elements werden anschließend gemittelt und bestimmen somit den Ausgang des *i*-ten ***A**verage **L**ocalized **S**ynchrony **D**etector*:

$$ALSD_i = \frac{1}{N} \sum_{k=i-N_1}^{i+N_2} GSD_i(f_k(t)) \qquad (3.4.6)$$

Die Formanten gehören zu den energiereichen Signalabschnitten, sie erregen daher oft auch benachbarte Filter. Individuelle Harmonische dagegen - welche nur gelegentlich in einem der Filterbänder auftreten - werden im Gegensatz zu den Formanten durch diese Methode unterdrückt. D.h. neben der Schärfung der Formanten kann zusätzlich eine Glättung bezüglich kurzzeitig auftretender spektraler Spitzen beobachtet werden.

Ebenfalls durch dieses Modell motiviert ist die von *[Kajita-96]* vorgestellte ***S**ub**B**and **COR**relation* -Analyse zur Extraktion robuster Merkmalvektoren.

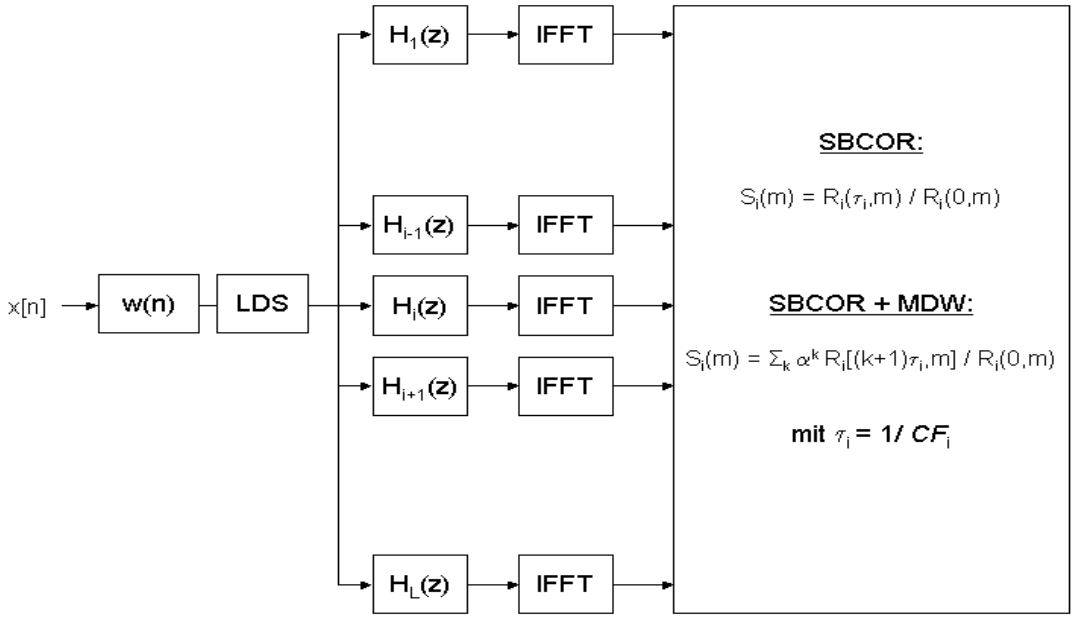

Abbildung 3.4.2 ***Die SBCOR Analyse (insbesondere deren Multi-Delay-Weighting Erweiterung) verzichtet vollständig auf dynamische Aspekte zeigt aber, dass diese Analysemethode ihrem Wesen nach ebenfalls durch ein LIN beschrieben werden kann.***

Zur Darstellung eines robusten Spektrums wurde das Eingangssignal zunächst in den Frequenzbereich transformiert und dessen Leistungsdichtespektrum mit dem Betragsquadrat der Übertragungsfunktionen einer Bark skalierten Filterbank multipliziert. Die zugehörige Teilband- Autokorrelationsfunktionen erhält man durch Anwendung der *IFFT*:

$$R_i(\tau, m) = \int_{-\infty}^{\infty} |H_i(f)|^2 X(f, m) \cos 2\pi f \tau \, df \qquad (3.4.7)$$

Unter Verwendung den Teilbandperiodizitäten $\tau_i = CF_i^{-1}$, werden dann folgende Merkmale berechnet:

$$S_i(m) = \frac{R_i(\tau_i, m)}{R_i(0, m)} \tag{3.4.8}$$

Derartige Merkmale zeigen bei weißen additiven Rauschstörungen erwartungsgemäß die gewünschte Robustheit, da die *AKF* der Störung in diesem Fall nur an der Stelle $\tau = 0$ signifikante Werte aufweisen kann. An der Stelle der inversen Mittenfrequenz dagegen ist der Beitrag der Stör- *AKF* vernachlässigbar. Weitaus interessanter ist dagegen eine Ansatzerweiterung hinsichtlich der Verwendung von Vielfachen der inversen Mittenfrequenz *[Kajita-98]*. Hier konnten die Autoren auf analytischem Wege nachweisen, dass mit der *AKF*-Methode im Grunde ebenfalls ein *LIN* nachgebildet wird, welches eine Schärfung des *SBCOR*- Spektrums zur Folge hat.

$$S_i(m) = \frac{1}{A}\sum_{k=0}^{\infty} \alpha^k R_i^m((k+1), \tau_i) / R_i^m(0)\ , \quad mit \quad A = \sum_{k=0}^{\infty} \alpha^k \tag{3.4.9}$$

Dabei konnte gezeigt werden, dass die berechnete Gewichtsfunktion (3.4.10) für den *i*-ten Kanal im Wesentlichen von der Mittenfrequenz und der dominanten Frequenz in diesem Kanal abhängt.

$$W_i(f) = \frac{(1-\alpha)(\cos 2\pi f_d^i / f_c^i - \alpha)}{1 - 2\alpha \cos 2\pi f_d^i / f_c^i + \alpha^2} \tag{3.4.10}$$

Die Gewichtsfunktion bildet demnach ein *LIN* nach, das aktuelle Gewicht wird in jedem Kanal durch das Verhältnis von dominanter Frequenz zur Mittenfrequenz bestimmt. Mit dem Parameter α kann der Grad der Schärfung des Spektrums eingestellt werden. Setzt man den Parameter auf Null, geht der Ausdruck (3.4.9) in das zuvor beschriebene *SBCOR*-Spektrum über. Auf die Nachbildung weiterer auditiver Eigenschaften wurde leider verzichtet. Von besonderer Wichtigkeit ist die festgestellte Äquivalenz der *LIN*- und der *AKF*-Methode, beide Methoden führen letztlich zur Schärfung des auditiven Spektrums. Der Forderung nach einer besonders robusten Darstellung genügt theoretisch die *ALSD*-Formulierung. In der Praxis findet man die Störquelle weißes Rauschen allerdings eher selten. Bei allen Untersuchungen mit dem Seneff-Modell oder deren Abwandlungen mussten daher bei andersartigen Störungen Performanzverluste hingenommen werden. Aus diesem Grund scheint es für Haarzellen-Modelle wünschenswert zu sein, ein Verfahren zur Rauschminderung zu verwenden, welches ein selbstnormalisiertes Störspektrum erzeugt.

3.5 PAS Modell nach O. Gitzha

Das ***E**nsemble **I**nterval **H**istogram [Gitzha-94]* besteht aus einer gehörrichtigen Filterbank mit einem nachfolgenden nichtlinearen Prozessor, welcher die Ausgänge der Filter in entsprechende neuronale Feuermuster umsetzt. Der nichtlineare Prozessor wird durch ein Array von Schwellwertdetektoren und Zählern gebildet. Die Intensitätsinformation des Stimulus wird durch die Anzahl der überschrittenen Schwellwerte berücksichtigt. Im Modell wurden die Schwellwerte logarithmisch verteilt. Die Frequenzinformation wird durch die Intervalle zwischen aufeinander folgenden Schwellwertdurchgängen positiver Steigung kodiert, somit wird die direktionale Auslenkung der Haare auf der Basilarmembran berücksichtigt. Das *EIH* verwendet ausschließlich Frequenz- und Intensitätsinformationen zur Modellierung neuronaler Feueraktivität, dynamische Eigenschaften der Haarzelle werden in diesem Modell nicht berücksichtigt.

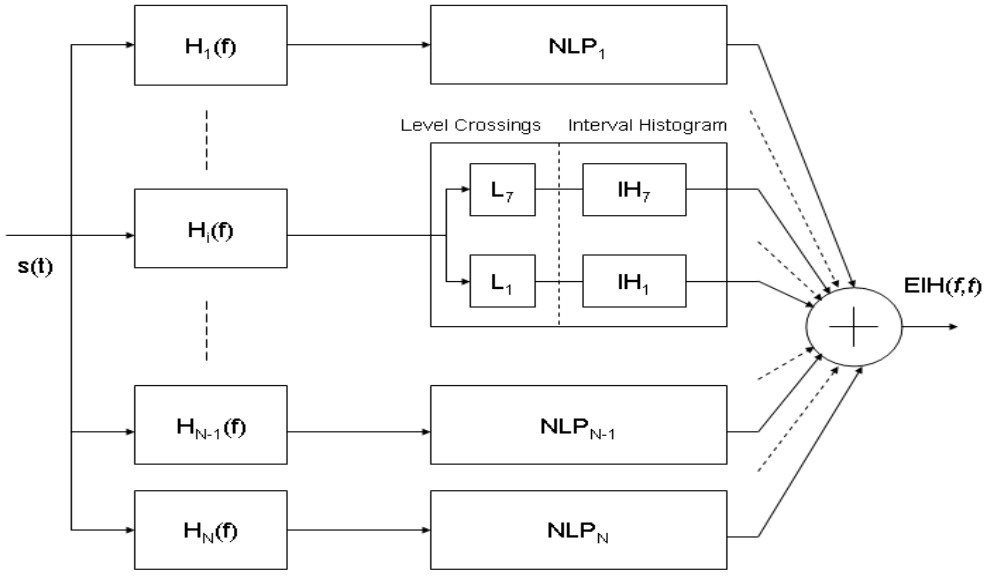

Abbildung 3.5.1: ***Das Blockschaltbild des EIH- Modells zeigt eine gehörrichtige Filterbank, gefolgt von einem Feld von Schwellwertdetektoren (Level Crossings) und Zählern (Interval Histogram), welches einen nichtlinearen Prozessor darstellt. Dieses Modell stellt ebenfalls die Umwandlung der mechanischen Bewegungsenergie entlang der Basilarmembran in neuronale Aktivität der Nervenfasern dar. Die direktionale Auslenkung der Haarzellen wird durch die Auswertung der Schwellwertdurchgänge mit ausschließlich positiver Steigung berücksichtigt. Dynamische Aspekte und Reduktion der Synchronität werden in diesem Modell nicht berücksichtigt.***

Die Realisierung des *EIH* zeigt eine interessante Parallele zur *ALSD*-Formulierung aus Abschnitt 3.4. Durch die Überlappung der Übertragungsfunktionen der Filterbank kann ein Teilband-Histogramm auch Frequenzkomponenten detektieren, die eigentlich den Nachbarkanälen zugeordnet werden sollten. Die nachfolgende Mittelung der Teilband-Histogramme hat zur Folge, dass solche Frequenzkomponenten - welche mehrere benachbarte Filter erregen können - mehrfach gezählt werden und somit verstärkt werden. Im Gegensatz dazu werden individuelle Harmonische die in den einzelnen Filtern nur sporadisch auftreten nur jeweils einfach gezählt und somit unterdrückt. Dies führt wie bei der *ALSD*-Formulierung zu einer robusten Darstellung des *EIH*-Spektrums.

In *[Kim-99]* wurde gezeigt, dass die Anordnung der Schwellwertdetektoren (d.h. die Werte der verschiedenen Schwellwerte) eine wichtige Rolle spielt.

Bei Störungen welche durch eine weiße Störquelle verursacht werden, haben die Autoren analytisch gezeigt, dass die Varianz der Intervalle von Schwellwertdurchgängen dann minimal ist, wenn sich der Schwellwert auf der Null-Linie befindet, hier wird dann naturgemäß von Nulldurchgangsintervallen gesprochen. In diesem Fall kann die Frequenz eines gestörten sinusförmigen Signals mit folgendem *SNR* gemessen werden:

$$SNR_i = \frac{2\pi / \omega_i}{\sigma_i^2} \quad \text{, mit} \quad \sigma_i^2 = E\{(t_k - t_{k-1})^2\} \tag{3.5.1}$$

Die t_k bzw. t_{k-1} bestimmen die Position aufeinander folgender Nulldurchgänge im Teilband i. Das bedeutet, dass eine robuste Berechnung der Teilband-Histogramme mit nur einer einzigen Schwelle möglich sein sollte. Obwohl analytisch gezeigt werden konnte, dass die Frequenzmessung bei höheren Schwellwerten empfindlicher gegenüber Rauschen reagiert, zeigte sich das *EIH* mit logarithmisch verteilten Schwellwerten gegenüber dem *EIH* mit dem Schwellwert 0 überlegen. Dies kann dadurch erklärt werden, dass die Intensitätsinformation im letzteren Fall nicht verwendet wurde. Aus diesem Grund haben die Autoren eine Methode vorgeschlagen, bei welcher der Maximalwert zwischen zwei positiven Nulldurchgängen logarithmisch in die Berechnung eingeht.

$$y(t,i) = \sum_{channel} \sum_{k=1}^{K} \delta_{i,j_k} \log(A_k), \quad 1 \le i \le N, \tag{3.5.2}$$

Hier wird mit N die Anzahl der Teilbänder und mit K die Anzahl der Schwellwertebenen bezeichnet. Die publizierten Erkennexperimente mit dieser Methode zeigten gegenüber dem klassischen *EIH*-Ansatz über den gesamten *SNR*-Bereich bessere Resultate. Mit dieser Methode liegt demnach auch eine robuste Methode vor, um die dominante Frequenz eines Teilbandes im Zeitbereich schätzen zu können. Auf diese Methode wird in Kapitel 7 zurückgekommen, dort wird ein *LIN* im Zeitbereich vorgeschlagen, das unter Verwendung der dominanten Frequenz realisiert wird.

3.6 Das Zentrale Auditorische System (CAS)

Die neuronale Architektur des Hörens beinhaltet sowohl das Periphere Auditorische System als auch das Zentrale Auditorische System (*engl. **C**entral **A**uditory **S**ystem*). Im Hinblick auf die technisch umsetzbaren Wirkprinzipien des Zentralen Auditorischen Systems werden hier nur 2 Kerngebiete der zentralen Hörbahn beschrieben, der *Colliculus Inferior* und der *Auditorische Cortex*.

An Größe übertrifft der *Colliculus Inferior* alle anderen Kerngebiete der zentralen Hörbahn, er ist das Hauptziel aller aufsteigenden Axone. Da die aufsteigenden Axone aus mehreren Kerngebieten der zentralen Hörbahn entstammen, die jeweils unterschiedliche Anteile des akustischen Signals kodieren, tragen sie sehr unterschiedliche Informationen. Zudem ist eine unterschiedlich große Zahl von Synapsen zwischengeschaltet, wodurch sich auch eine zeitliche Staffelung der eingehenden Erregung ergibt. Aufgrund dieser strukturellen Eigenarten des *Colliculus Inferior* liegt zunächst die Vermutung nahe, dass er eine entscheidende Rolle beim binauralen Hören und damit bei der Lokalisation von Schallquellen spielt.

Der *Colliculus Inferior* wird demnach als Schaltzentrale für eingehende Erregungen betrachtet, diese müssen aber nicht unbedingt nur den akustischen Nervenbahnen zugeordnet sein, sie können auch anderen Sinnesmodalitäten entstammen. So gibt es einige Hinweise darauf, dass im Falle von Mehrdeutigkeiten Informationen aus den visuellen Nervenbahnen nicht nur zur Bildung von Hörempfindungen herangezogen werden, sondern dass die visuellen Informationen die akustischen Informationen sogar dominieren können *[Green-91]*. Demnach könnte die Bildung von Hörempfindungen auf der multimodalen Verarbeitung von raumzeitlichen Ereignissen (akustisch, visuell) und der anschließenden Integration der zugehörigen spektrotemporalen Merkmale zu einem in sich konsistenten Klangbild basieren.

Vom *Auditorischen Cortex* dagegen nimmt man an, dass dort eine Schärfung des Klangbilds stattfindet und zusammengehörige Merkmale gruppiert werden. Dieses Kerngebiet der zentralen Hörbahn übernimmt also die Organisation der Merkmale und hat die Bildung auditiver Objekte zur Folge. An dieser Stelle sei noch einmal darauf hingewiesen, dass auch in diesem Kerngebiet aufsteigende und absteigende Nervenbahnen zu finden sind.

Die Eigenschaften dieser beiden Teile des Zentralen Auditorischen Systems bilden die Grundlage für die Theorie der *Auditory Scene Analysis*. Sie beschreibt die kognitiven Aufgaben, die das Zentrale Auditorische Systems bei der Identifikation und Bildung von auditorischen Objekten in komplexen akustischen Situationen zu lösen hat. Dazu gehört, dass das Zentrale Auditorische System erkennen muss, wie viele Schallquellen in der Umgebung aktiv sind, woher sie kommen und ob bedeutungsvolle Informationen übertragen werden.

Da ein recheneffizientes auditives Modell neben den wesentlichen Eigenschaften des *PAS* nun auch Eigenschaften des *CAS* in sich tragen sollte, werden nun einige technisch umsetzbare Eigenschaften des *CAS* hervorgehoben.

Eine dieser Eigenschaften ist die bereits in der Einführung erwähnte Bindungseigenschaft. Unter diesem Begriff wird die Fähigkeit verstanden, zusammengehörige Signalkomponenten eines auditiven Objekts hervorzuheben und vom Hintergrund zu trennen. Eine wichtige Voraussetzung zur Bildung von auditiven Objekten besteht zunächst darin, zusammengehörige Signalkomponenten identifizieren zu können. In der einschlägigen Fachliteratur sind hierzu einige Vorschläge beschrieben worden, so könnte man sich bspw. auf diejenigen Signalkomponenten konzentrieren, welche gleichzeitig einsetzen und sich anschließend relativ langsam und kohärent entwickeln (z.B. die Harmonischen der Sprachgrundfrequenz). Als möglicher Indikator für die Zusammengehörigkeit von Signalkomponenten kann die gemeinsame Einhüllende von Teilbandsignalen betrachtet werden *[Hasegawa-97]*.

Deren Ursprung ist wiederum auf die Sprachgrundfrequenz zurückzuführen. Werden nun diejenigen Signalkomponenten, welche sich durch eine gemeinsame Einhüllende auszeichnen besonders betont, dann führt dies auf so genannte pitchkohärente Merkmale (*engl.* ***P****itch* ***C****oherent* ***F****eatures*).

Mit pitchkohärenten Merkmalen ist nun einerseits die Vorstellung verbunden, dass diese in stimmhaften Abschnitten zu einer Verbesserung des SNR führen und andererseits die Bildung von auditiven Objekten unterstützt. Eine wichtige Voraussetzung für die Bildung von pitchkohärenten Merkmalen ist die Bestimmung der Sprachgrundfrequenz, bei Störungen oder einer Verletzung des Redundanzprinzips können hier schnell Sprünge gemessen werden, die dann wiederum das Kontinuitätsprinzip verletzen. Um derartige Probleme zu vermeiden kann auf ein weiteres bekanntes Prinzip zurückgegriffen werden: Variation, Selektion und Reproduktion. Die technische Umsetzung dieses Prinzips ähnelt im Kern einem evolutionären orientierten Algorithmus und dient vor allem der Auflösung von Mehrdeutigkeiten. Ein wichtiges Element ist dabei die Rückführung von Vorhersagen über erwartete Eingangssignale. Im Folgenden wird gezeigt, dass dieses Prinzip durchaus mit den bisher beschriebenen Gehirnstrukturen in Einklang stehen kann.

Wie bereits mehrfach erwähnt, findet der Informationsfluss von den Sinneszellen über das Periphere Auditorische System bis hin zum Zentralen Auditorischen System sowohl auf afferenten als auch auf efferenten Nervenbahnen statt. Untersuchungen an Tieren haben gezeigt, dass sich insbesondere die Erkennung von vokalischen Lauten deutlich verschlechtert, wenn die efferenten Nervenbahnen unterbrochen werden *[Dusan-05]*. Zudem ist aus der Hirnforschung bekannt, dass ein inneres Abbild der motorischen Strukturen und deren Dynamik ausgenutzt werden kann, um Eingangssignale vorherzusagen. Nun kann man vermuten, dass auf den efferenten Bahnen Vorhersagen bezüglich des zu erwartenden Eingangssignals zurückgeführt werden und dass diese Vorhersagen in stimmhaften Abschnitten stärker berücksichtigt werden als in stimmlosen Abschnitten. Diese (hypothetischen) Eigenschaften des *CAS* sollten demnach zusätzlich in die auditive Modellierung aufgenommen werden. Die Wirksamkeit der Modellierung dieser zusätzlichen Eigenschaften muss dann unter Berücksichtigung der zur Verfügung stehenden Ressourcen durch das Experiment überprüft werden.

3.7 Elementare Eigenschaften auditiver Modelle

Eine Zusammenfassung derjenigen Eigenschaften welche ein echtzeitfähiges auditives Modell in sich vereinen sollte und konsistent zur Artikulationstheorie sind, werden durch die folgenden Aufzählungen gegeben.

Mit einem *PAS*- Modell sollten die nachfolgenden auditiven Eigenschaften erfasst werden:

- Bark- oder Mel-Skalen Filterung,
- Kompressionskennlinie, Intensitäts-Lautheits-Transformation
- Kurzzeit-Adaption
- Maskierung (Zeitbereich)
- Laterales Inhibitions Netzwerk (*LIN*)

Die benötigten *CAS*- Eigenschaften können durch folgende Mechanismen nachgebildet werden:

- Merkmalsbindung
- Variation von Merkmalen
- Vorhersage von Merkmalen
- Selektion von Merkmalen, Auflösung von Mehrdeutigkeiten

4 Auditive Modelle mit PAS Eigenschaften

In diesem Abschnitt werden nun Multibandmodelle vorgestellt, bei denen das Haarzellen-Modell von *[Martens-90]* Anwendung findet. Dieses Modell ist insbesondere hinsichtlich des Implementierungsaufwands eine interessante Alternative zum Modell von S. Seneff. In Abschnitt 4.1 wird zunächst ein Multiband-Modell im Zeitbereich vorgestellt, d.h. sowohl die Filterbank als auch die Haarzellen Modelle werden mit der vollen Abtastrate betrieben *[Vereecken-95]*. Bevor auf dieses Modell eingegangen wird, zeigt eine Systemanalyse des Haarzellen-Modells dessen wesentliche Übertragungseigenschaften.

Bei dem zweiten Modell in Abschnitt 4.2 handelt es sich um ein Multiband-Modell im Frequenzbereich *[Perdigao-98]*. Das Eingangssignal wird einer *STFT*-Analyse unterzogen und im Frequenzbereich mittels einer Gammatone-Filterbank in Teilbänder zerlegt. Die Haarzellen-Modelle werden anschließend mit der Framerate betrieben. Hinsichtlich der Robustheit der Modelle werden in Kapitel 6 Verfahren zur Rauschunterdrückung vorgestellt, welche die Methode der Spektralen Subtraktion mit der Maskierungstechnik kombinieren, um den dynamischen Bereich des Haarzellenmodells dem Nutzsignal zur Verfügung zu stellen.

Diese beiden Verfahren unterscheiden sich vor allem dadurch, dass sie im Zeit- bzw. Frequenzbereich angesiedelt sind. Vorteilhaft bei dem Modell im Frequenzbereich ist der deutlich geringere Rechenaufwand. Allerdings können zeitlich lokale Eigenschaften des Signals im Frequenzbereich lediglich durch globale Basisfunktionen approximiert werden. Ebenso werden lokal wirksame Störungen bei der Analyse auf das gesamte Spektrum verteilt. In Abschnitt 4.3 wird daher ein Modell vorgeschlagen, welches diesen Nachteil durch den Einsatz eines mehrfach auflösenden Verfahrens überwindet. Eine effiziente Implementierung auf der Basis einer kritisch abgetasteten Filterbank gestattet es, zwei Ansätze zu verfolgen, welche die Verwendung des Haarzellen Modell ermöglichen: Zum einen können die Haarzellen-Modelle mit den Kurzzeit-Effektivwerten der Teilbänder gespeist werden, zum anderen können die Haarzellen-Modelle auch direkt mit den Teilbandsignalen mit jeweils unterschiedlichen Abtastraten betrieben werden. Beide Ansätze erfordern nur geringe Modifikationen hinsichtlich der Verwendung der in Kapitel 6 beschriebenen Verfahren zur Rauschminderung. Da die bisher besprochenen Modelle kein *LIN* verwenden, werden abschließend Möglichkeiten zur Implementierung der *LIN*- Funktionalität zum Zwecke der spektralen Schärfung des auditiven Spektrums vorgeschlagen.

4.1 Auditives Modell mit PAS Eigenschaften im Zeitbereich

Dieses Haarzellen Modell wird ebenfalls in dem später beschriebenen Modell im Frequenzbereich eingesetzt, hier wird zunächst eine Rauschunterdrückung im Zeitbereich verwendet, welches im Mittel ein konstantes flaches Störspektrum erzeugt. Das Haarzellenmodell wurde von *[Martens-90]* vorgestellt. Für die Berechnung der mittleren Feuerrate geben die Autoren folgende Gleichung an:

$$\mu_f = \frac{f_{sat}\ \mu_V}{\left(B + \sqrt{q}\right)^2} \tag{4.1}$$

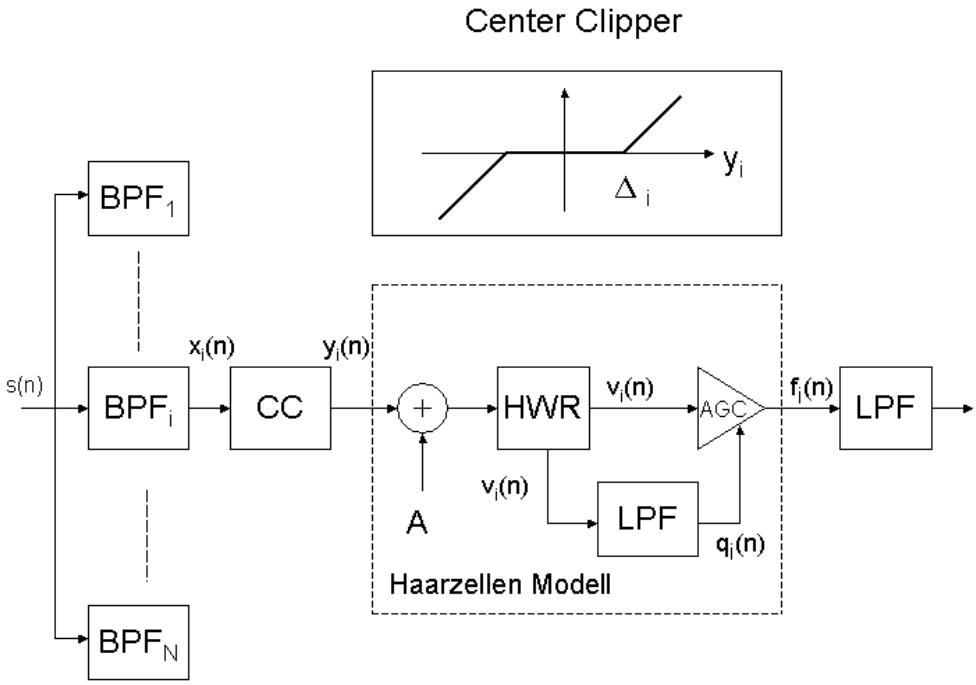

Abbildung 4.1.1 ***Das vereinfachte Haarzellenmodell nach Martens dient als Basis für die integrierte lineare bzw. nichtlineare Rauschunterdrückung. Dabei wird der Schwellwert A in Abhängigkeit vom Rauschniveau adaptiert, so dass der dynamische Bereich erhalten bleibt.***

Eine Analyse des Übertragungsverhaltens des *HWR*-Blocks zeigt, dass der Mittelwert μ_v als Funktion des Effektivwerts σ_Y beschrieben werden kann. Das Verhalten der Funktion wird im Wesentlichen durch die Kompressionseigenschaft charakterisiert. Untersuchungen zur Anregung mit weißen Rauschen (σ = 5 dB ...80 dB) zeigen im doppelt logarithmischen Maßstab einen fast linearen Zusammenhang zwischen dem Effektivwert des Eingangs und dem Mittelwert des *HWR*-Ausgangs.

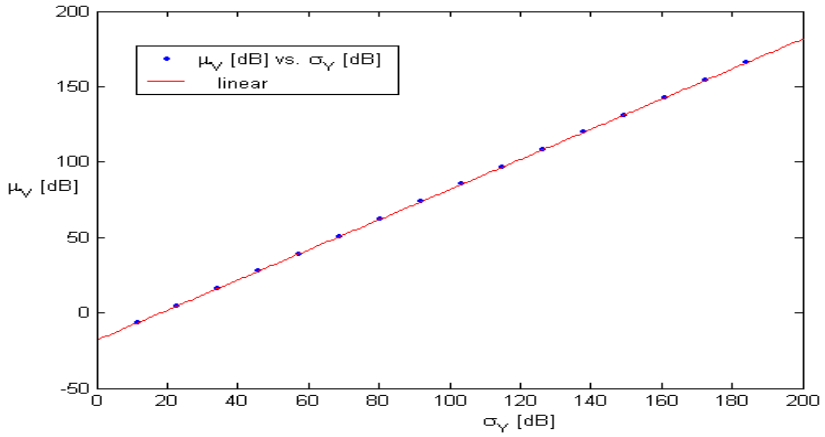

Abbildung 4.1.2 ***Das Übertragungsverhalten der Halbwellengleichrichtung kann im doppelt logarithmischen Maßstab durch eine lineare Funktion angenähert werden.***

Das dem *HWR*-Block folgende *AGC*-Modell legt den dynamischen Bereich der Amplituden der Filterbankausgänge fest. Unterhalb einer Schwelle von 20 dB werden die Amplituden auf die spontane Feuerrate f_{spo} abgebildet. Die sigmoide Kennlinie verläuft innerhalb des dynamischen Bereichs fast linear. Bei starken Erregungen wird die Feuerrate auf den Wert f_{sat} begrenzt.

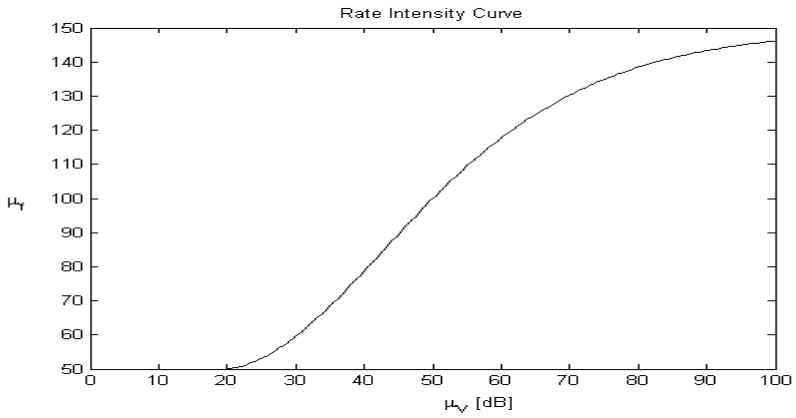

Abbildung 4.1.3 Die Übertragungskennlinie des Haarzellenmodells entspricht einer Sigmoidalfunktion, der dynamische Bereich beträgt etwa 80 dB.

Da der Erwartungswert μ_v im doppelt logarithmischen Maßstab fast linear vom Effektivwert σ_Y des Eingangssignals abhängt, kann die mittlere Feuerrate μ_f im Prinzip auch als Funktion von $\log\sigma_Y$ beschrieben werden.

4.2 Auditives Modell mit PAS Eigenschaften im Frequenzbereich

[Perdigao-98] berichtet über den erfolgreichen Versuch, ein recheneffizientes auditives Modell im Frequenzbereich zu realisieren. Dieses konnte es mit dem *MFCC*-Verfahren bezüglich des Rechenaufwands durchaus aufnehmen. Die Motivation für diesen Schritt war zum einen die in der Literatur beschriebene Überlegenheit von auditiven Modellen gegenüber den konventionellen Verfahren bei Spracherkennungsproblemen, zum anderen die Anforderungen aus der Praxis bezüglich Rechenzeit und Speicherplatz zu erfüllen. F. Perdigao zeigte auch im Hinblick auf Robustheitseigenschaften, dass das Haarzellenmodell hier ebenfalls ein gewisses Potential aufzuweisen hat. Darüber hinaus konnte durch die Analyse des verwendeten Modells und dem Vergleich mit anderen Signalverarbeitungsmethoden ein tieferes Verständnis für die grundlegenden Mechanismen der peripheren auditiven Verarbeitung gewonnen werden. Die Hauptcharakteristiken auditiver Verarbeitungsmethoden sind in Abbildung 4.2.1 ersichtlich.

Aus Rechenzeitgründen wird von Perdigao ebenfalls das Haarzellen Modell nach *[Martens-90]* verwendet. Es kann - wie oben beschrieben - durch eine nichtlineare Kompressionskennlinie charakterisiert werden. Diese ist durch eine Erregungsschwelle, einer Schwelle für die spontane Feuerrate und einer Schwelle für den Sättigungspunkt der Feuerrate gekennzeichnet. Der abschließende Tiefpass modelliert die Synchronitätsreduktion.

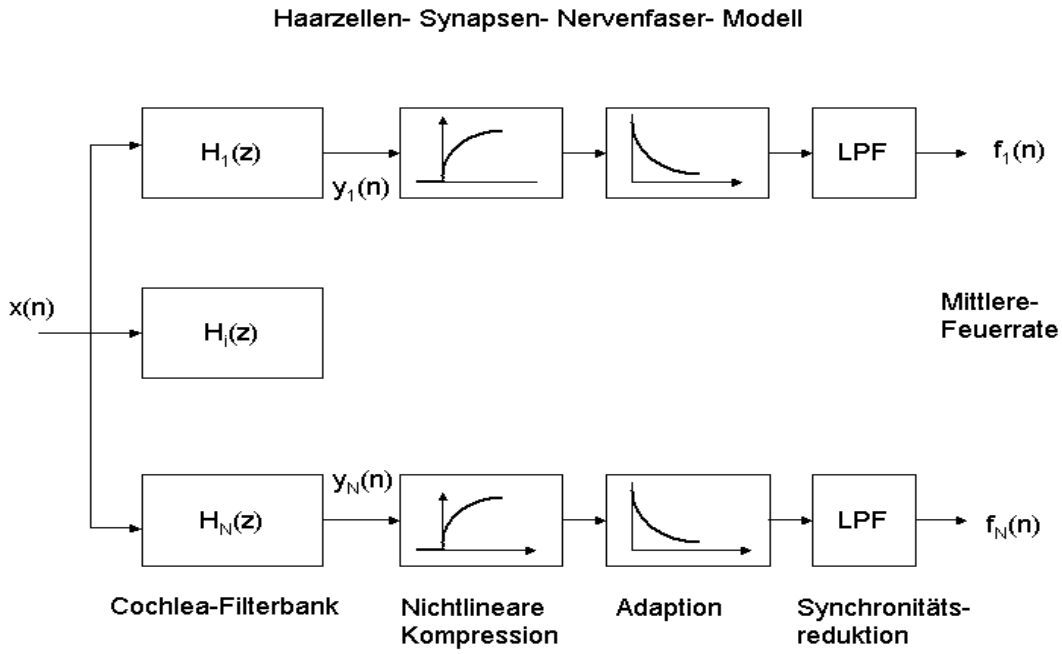

Abbildung 4.2.1 ***Modell des Peripheren Auditiven Systems nach F. Perdigao. Nach der Filterbank folgen die die Elemente Nichtlineare Kompression, Adaption und Synchronitätsreduktion.***

Dank der dynamischen Eigenschaften sollte das Modell in der Lage sein, starke und schnell veränderliche Signalkomponenten zu verstärken. Dies gilt nicht nur für sprunghafte Segmente sondern auch für schnelle Änderungen in den Formantfrequenzen. Erste Experimente zur Robustheit vergleichen das Antwortverhalten der Haarzellenmodelle mit den logarithmierten Filterbankenergien bezüglich eines Chirpsignals (von 2 kHz nach 0.5 kHz in 200 ms) *[Perdigao-97]*. Die additive Überlagerung von Rauschen führt in beiden Fällen zu einer Einschränkung des dynamischen Bereichs. Bei schnell veränderlichen Signalen kann aber im Vergleich zum Antwortverhalten der logarithmierten Filterbank-energien ein besseres lokales SNR beobachtet werden. Die logarithmierten Filterbank-energien werden zum einen fast vollständig vom Rauschen maskiert, zum anderen ist die Antwortkurve wesentlich breiter als die Antwortkurve des Haarzellenmodells.

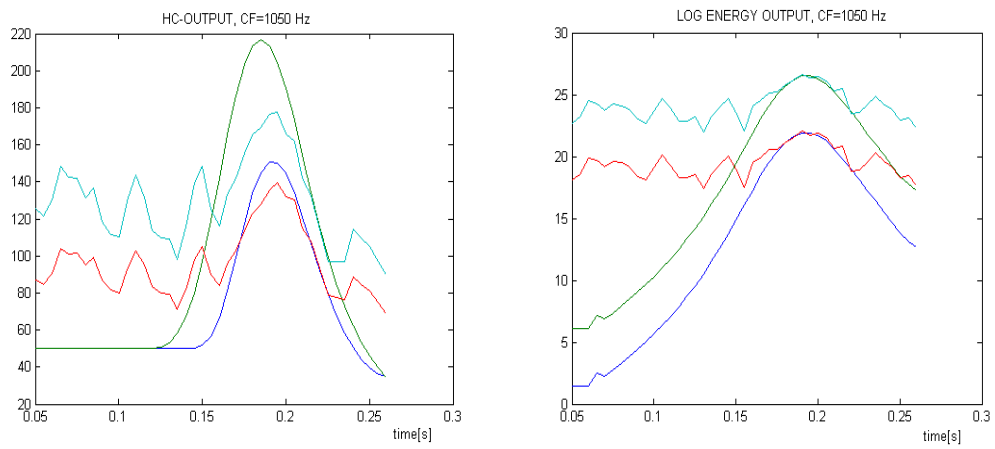

Abbildung 4.2.2: ***Experimente mit einem Chirp- Signal, die glatten Kurven entsprechen einem SNR = ∞, und Amplituden von V = 60 dB bzw. V = 80 dB. Die anderen Kurven wurden für SNR von 20 dB und V= 60 dB bzw. bei einem SNR von 0dB und V = 80 dB gemessen.***

Ein Vergleich mit dem *lin-log-RASTA*- Verfahren führte zur Schlussfolgerung, dass das Adaptionsprinzip durch eine Kompressionsstufe gefolgt von einem linearen Filter erklärt werden kann *[Perdigao-98]*. Neben der Ausnutzung von Modulationseffekten wie bei dem *lin-log-RASTA*- Verfahren findet man hier ebenfalls die Anwendung der Maskierungs-Technik.

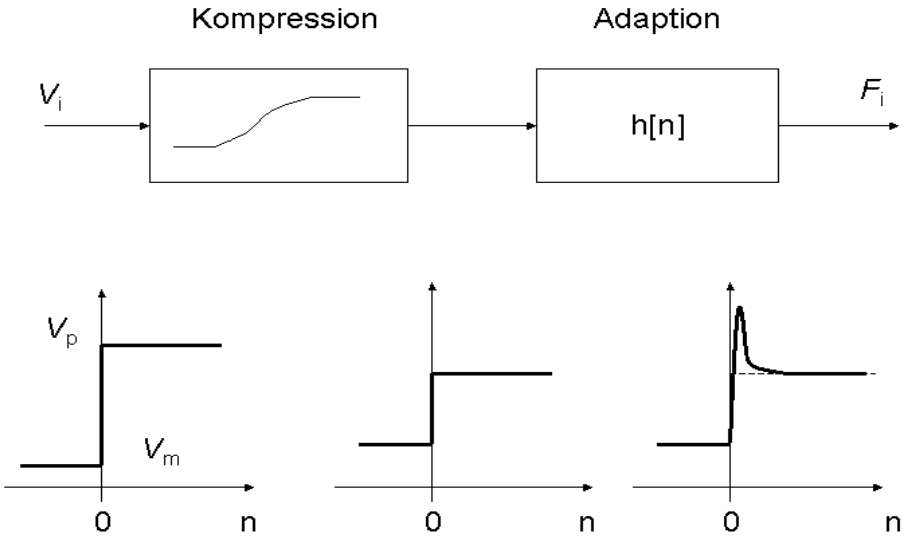

Abbildung 4.2.3: ***Modell des Adaptionsverhaltens. Das Antwortverhalten auf einen Sinus-Burst der Amplitude V_p kann durch eine nichtlineare Kompression gefolgt von einem linearen Filter modelliert werden.***

Nachteile / Probleme:

- Die nichtlinearen Komponenten in Abbildung 4.2.1 fordern zunächst eine Realisierung mit der vollen Abtastrate.
- Die Verschiebung des dynamischen Bereichs durch Rauscheinflüsse fordert eine Adaption der Erregungsschwelle

In einem nächsten Schritt wurde untersucht, ob man die Adaptionscharakteristik auch dann beobachten kann, wenn als Eingangsignal Kurzzeit-Effektivwerte verwendet werden *[Perdigao-99]*. Diese Idee kann durch die Beobachtungen bezüglich des *HWR*-Übertragungsverhaltens begründet werden. Wenn der Mittelwert annähernd linear mit dem Effektivwert zusammenhängt, dann bleibt die Einhüllende des Effektivwertes bei der *HWR*-Übertragung ebenfalls erhalten. Durch Variation der Filterkoeffizienten im Martens-Modell konnten durch einen Vergleich mit der Adaptionscharakteristik bei voller Abtastrate diejenigen Modellparameter gefunden werden, mit denen sich die Adaption auf der Basis von Kurzzeit-Effektivwerten vernünftig modellieren lässt. Berücksichtigt man zudem, dass der Effektivwert der einzelnen Teilbänder mit der Parsevalschen Beziehung auch im Frequenzbereich ermittelt werden kann,

$$\sum y^2(n) = \frac{1}{N}\sum |Y(k)|^2 \qquad (4.2.1)$$

so folgt daraus die Berechnung der Leistung der Teilbandsignale nach (4.2.2).

$$\mathbf{y}_i = \mathbf{H}_i \mathbf{x} \quad \text{bzw.} \quad y_i^{\,P} = \frac{1}{N} \sum_N y_i^{\,2}(n) \tag{4.2.2}$$

Mit (4.2.1) und (4.2.2) erhält man die Berechnungsvorschrift für den Effektivwert im Frequenzbereich:

$$y_i^{\,RMS} = \sqrt{y_i^{\,P}} = \frac{1}{N} \sqrt{\sum_N |\, Y_i\ (k)\,|^{\,2}} \tag{4.2.3}$$

Als Analysemethode liegt immer noch die *FFT* zugrunde, lokale Störeinflüsse werden demnach auf das gesamte Spektrum verteilt. Ebenso werden kurzzeitige Phänomene des Signals innerhalb des Analysefensters nur schlecht aufgelöst. Darüber hinaus muss auch noch das Problem der Adaption der Erregungsschwelle bei Rauscheinfluss gelöst werden. Dieses Problem wird in Kapitel 6 behandelt.

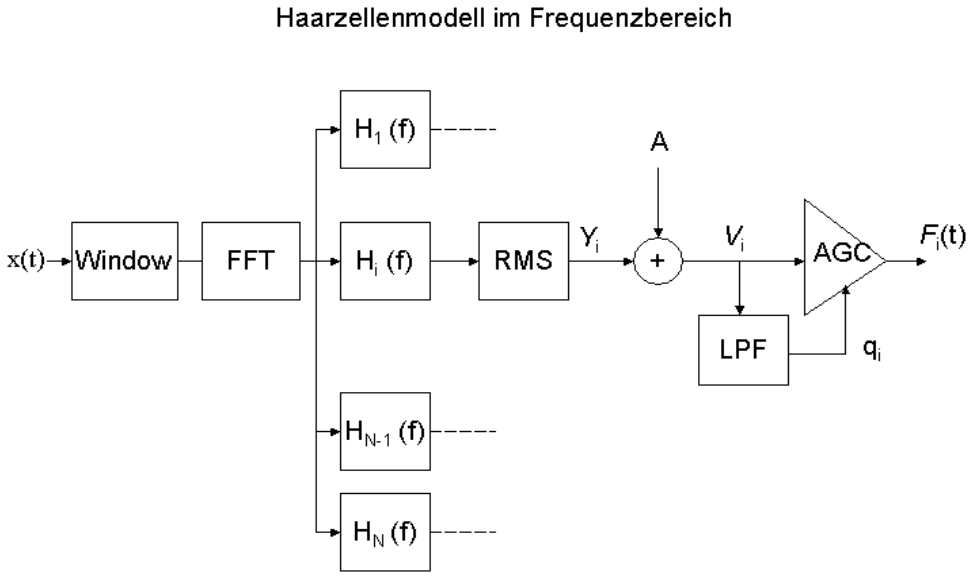

Abbildung 4.2.6: ***Modellierung der dynamischen Eigenschaften des auditorischen Systems im Frequenzbereich, hierbei kommt das Haarzellenmodell von [Martens-90] zur Anwendung. Der Kurzzeit-Effektivwert lässt sich ebenfalls leicht im Frequenzbereich berechnen.***

4.3 Auditives Modell mit PAS Eigenschaften im Wavelet-Bereich

Bei dem zuvor beschriebenen auditiven FFT-Modell wird die eingangseitige Signalanalyse von einer gefensterten Fouriertransformation geleistet. Diese in der Sprachverarbeitung übliche Transformation bietet wegen der festen Fensterbreite keine Flexibilität bezüglich der Auflösungseigenschaften im Zeit- bzw. Frequenzbereich, sie eignet sich daher eigentlich nur für Analyse stationärer periodischer Signale.

Für nichtstationäre Signale stellte die gefensterte Fourieranalyse einen vernünftigen Kompromiss dar. Mit der kontinuierlichen Wavelet-Transformation *CWT* aber auch anderen Zeit- Frequenz- Analysemethoden wie z.B. der Wigner-Verteilung können für nichtstationäre Signale fortschrittlichere Analysemethoden verwendet werden.

Bei diesen mehrfach auflösenden orthogonalen Transformationen werden die hohen Frequenzen eines Signals mit kurzen Wavelets und die tiefen Frequenzen mit längeren Wavelets analysiert. Da sich das Signal transformieren lässt ohne das man überhaupt auf die Waveletfunktionen zurückgreifen muss, sondern stattdessen die Transformation durch eine Faltung von Signal und Filtern ausführen kann, sind effektive Realisierungen mittels kritisch abtastenden Filterbänken möglich. Eine weitere Erhöhung des Freiheitsgrades bei der Analyse von nichtstationären Signalen wurde durch die Einführung der ***D**iskreten **W**avelet **P**acket **T**ransformation* erzielt. Diese Flexibilität gewinnt man durch die unabhängige Variation von Frequenz, Position und Fensterbreite der Wavelets.

Die *DWPT* zeigt eine Reihe von zusätzlichen Eigenschaften, die bisher in der Sprachverarbeitung recht wenig ausgenutzt wurden. So kann die orthogonale Darstellung bei bestimmten Signalen dadurch redundant werden, dass sich die Eigenschaften eines Koeffizienten aus denen benachbarter Koeffizienten entnehmen lassen. Derartige Korrelationen sind dann durch das Signal selbst bedingt und können sich über mehrere Skalen hinweg ausbreiten.

Eine weitere interessante Eigenschaft ist die, dass zu einem gegebenen Zeitpunkt immer nur sehr wenige Wavelet- Koeffizienten aktiv sind. Welche Koeffizienten gerade aktiv sind, ändert sich von Zeit zu Zeit und hängt vom aktuellen Signalzustand ab. Dies nennt man eine spärlich verteilte Transformation. Im Gegensatz zu kompakten effizienten Transformationen wie der ***D**iskreten **C**osinus **T**ransformation* oder der ***P**rincipal **C**omponent **A**nalysis* (hier wird die Signalinformation auf nur wenige Koeffizienten reduziert), zerlegt die *DWPT* das Signal in nur *wenige aktive* Koeffizienten. Die Information über das Signal wird also über viele aber nicht gleichzeitig aktive Koeffizienten verteilt. Um diese Information nutzen zu können, wird ein Verfahren benötigt das auf einer kleinen aber zu jedem Zeitpunkt verschiedenen Anzahl von Koeffizienten basiert. Solche Verfahren könnten bspw. auf dem Prinzip eines Assoziationsspeichers beruhen, dieser Ansatz wird hier aber nicht weiter verfolgt.

4.3.1 Die Wavelet-Transformation

Neben der spektralen Analyse durch die *STFT* ermöglicht auch die *WT* eine spektrale Analyse von quasistationären Signalen. Im Gegensatz zur *STFT* werden aber bei der *WT* zur Approximation von Funktionen lokal begrenzte Basisfunktionen unterschiedlicher Ausdehnung verwendet, diese Basisfunktionen werden in der Literatur als Wavelets bezeichnet.

Wavelets sind eine Familie von Funktionen, welche durch Skalierung und Verschiebung aus einem lokal begrenzten Mutter- Wavelet gebildet werden.

$$\Psi_{s,\tau}(t) = \frac{1}{\sqrt{s}} \Psi\left(\frac{t-\tau}{s}\right) \tag{4.3.1}$$

Mit dem Parameter s kann das Mutter-Wavelet gedehnt oder gestreckt werden, so dass der Ausdehnungsbereich des Mutter-Wavelets modifiziert werden kann. Der konstante Faktor in (4.3.1) dient der Energienormalisierung. Mit dem Parameter τ kann das Mutter-Wavelet verschoben werden. Darüber hinaus muss das Mutter-Wavelet die Eigenschaft der Mittelwertfreiheit erfüllen:

$$\int \Psi(t)\, dt = 0 \tag{4.3.2}$$

Die Wavelet-Transformation einer Funktion *f(t)* erhält man durch Korrelation des Signals mit den verschiedenen Wavelets $\Psi_{s,\tau}(t)$.

$$WT_x(\tau, s) = \frac{1}{\sqrt{s}} \int f(t) \Psi^* \left(\frac{t-\tau}{s} \right) dt \qquad (4.3.3)$$

Solange für $|s| > 0$ und τ keine weiteren Einschränkungen gelten, bezeichnet man (4.3.3) als kontinuierliche Wavelet-Transformation (*CWT*).

Die Substitution $s = 2^{-j}$ und $\tau = k\,2^{-j} \quad k, m \in Z$ in (4.3.1) führt auf dyadische Wavelets.

$$\Psi_{j,k}(t) = \sqrt{2^j}\,\Psi(2^j t - k) \qquad (4.3.4)$$

Wendet man diese Familie von Basisfunktionen bei der Transformation eines Signals an, spricht man von der diskreten Wavelet-Transformation (*DWT*) oder auch von der schnellen Wavelet-Transformation (***F**ast **W**avelet **T**ransformation*).

Die *DWT* bzw. *FWT* kann als kritisch abgetastete Filterbank in Baumstruktur realisiert werden (siehe Anhang C). Eine besondere Bedeutung kommt dabei den orthogonalen bzw. biorthogonalen Filterbänken zu. Diese zeichnen sich durch eine besonders einfache und effiziente Implementierung aus.

4.3.2 Zum Einsatz der Wavelet-Transformation in der Sprachverarbeitung

Bereits seit einigen Jahren wird über Experimente zur Phonemerkennung mit der *WT* berichtet. Häufig werden zur Argumentation die guten Zeit- und Frequenzauflösungseigenschaften hervorgehoben. Ein weiteres Argument für den Einsatz der *WT* ist aber auch das folgende: ein Sprachsignal-Frame kann neben dem zentralen Phonem auch Informationen über benachbarte Phoneme enthalten. Wenn eines dieser Phoneme stimmhaft und das andere stimmlos ist, dann werden die tiefen Frequenzen durch das stimmhafte Phonem dominiert, während die hohen Frequenzen eher dem stimmlosen Phonem zuzuordnen sind. Hier liegt eine Möglichkeit vor, die Asynchronität von Phonemübergängen zu berücksichtigen und voneinander unabhängige Teilbänder zu verarbeiten.

In *[Fu-96]* wurden die Analysemethoden *CWT, DWT* und *STFT* in einem Experiment zur Phonemerkennung untersucht. Die *CWT* erwies sich gegenüber der *DWT* aufgrund der flexibleren Auflösung überlegen. Obwohl die *CWT* im Gegensatz zur *STFT* sowohl die Formanten als auch die Harmonischen gut auflösen konnte, war der Gewinn in der Erkennrate nur marginal. In beiden Fällen wurden die Merkmalvektoren auf der Basis von Mel-Cepstren berechnet und die Phoneme als HMM's mit drei Zuständen modelliert. Bessere Ergebnisse beim Einsatz der *WT* konnten bei der Kodierung, Pitchdetektion und Segmentierung von Sprache *[Fontaine-95]* erzielt werden.

Ermutigende Ergebnisse im Anwendungsbereich Spracherkennung konnten dagegen von *[Jabloun-95]* und später *[Kryze-99]* präsentiert werden. Die Autoren verwendeten eine Signalzerlegung basierend auf Wavelet-Paketen, welche als biorthogonale Filterbank mit Perfekter Rekonstruktion implementiert werden kann (Abb. 4.3.1). Die Zerlegung der Frequenzbänder wurde ähnlich der Mel-Skalen Zerlegung vorgenommen.

Wie in Anhang C.4 gezeigt wird, benötigt man zur Konstruktion einer solchen Wavelet-Filterbank ein Halbbandfilter, welches für eine spektrale Faktorisierung geeignet ist.

Ausgehend von der allgemeinen Darstellung einer Halbbandfilter- Übertragungsfunktion mit *4i-1* Koeffizienten

$$P_i(z) = \frac{1}{2}\sum_{n=1}^{i} p_i(2n-1)\,(z^{-2n+1} + z^{2n-1}) \tag{4.3.5}$$

werden die Koeffizienten nach der Lagrangschen Interpolationsformel berechnet:

$$p_i(2n-1) = \frac{(-1)^{n+i-1}\prod_{k=1}^{2i}(i-k+1/2)}{(i-n)!(i-1+n)!(2n-1)} \tag{4.3.6}$$

Das Lagrange Filter hat eine *2i*-fache Nullstelle bei *z=-1*, *i-1* Nullstellen innerhalb des Einheits-kreises und *i-1* dazu am Einheitskreis gespiegelte Nullstellen. Wegen der hohen Anzahl der Nullstellen bei *z=-1* erfüllt das Filter die Eigenschaft der Regularität und eignet sich daher besonders für den Entwurf von Wavelets. Die Übertragungsfunktion der zugehörigen Tief- bzw. Hochpassfilter erhält man nach *[Kim-92]*.

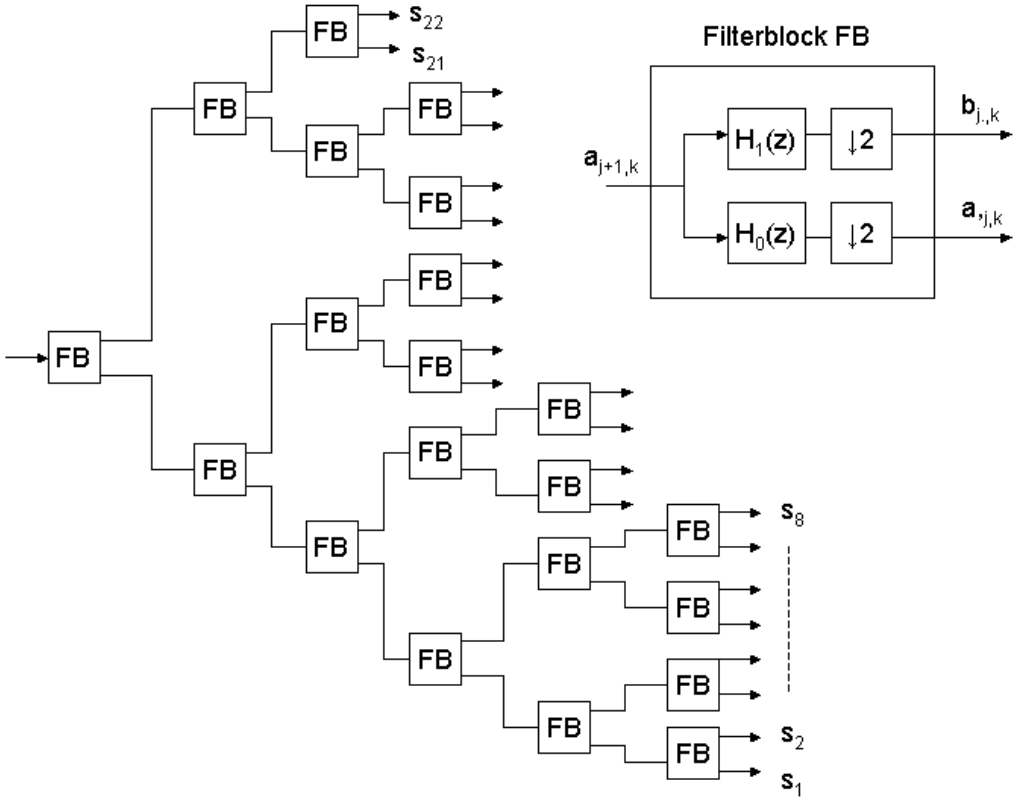

Abbildung 4.3.1: ***Wavelet-Paket-Zerlegung, die Frequenzbänder werden ähnlich der MEL- Skale zerlegt. Das Filter $H_0(z)$ liefert den Tiefpassanteil, $H_1(z)$ den Hochpassanteil. Nach der Unterabtastung decken beide Signale wieder das gesamte Nyquistintervall ab.***

In *[Jabloun-95]* wurden ausgehend von einem Halbbandfilter für den Analyseblock folgende Filter verwendet:

$$H_0(z) = \frac{1}{2} + \frac{9}{32}(z^1 + z^{-1}) - \frac{1}{32}(z^3 + z^{-3})$$

und (4.3.7)

$$H_1(z) = \frac{1}{2} - \frac{9}{32}(z^1 + z^{-1}) + \frac{1}{32}(z^3 + z^{-3})$$

Um in der Terminologie der Sprachsignalverarbeitung zu bleiben, bezeichnen wir das längste Mittelungsfenster $N_{l,max}$ als Fensterlänge, die Länge des kleinsten Fensters $N_{l,min}$ entspricht der Frame-Größe. Ähnlich wie bei den konventionellen Methoden werden demnach die Fenster entsprechender Größe (zentriert bezüglich der Position des kleinsten Fensters) über jedes Teilband geschoben (Abb. 4.3.2).

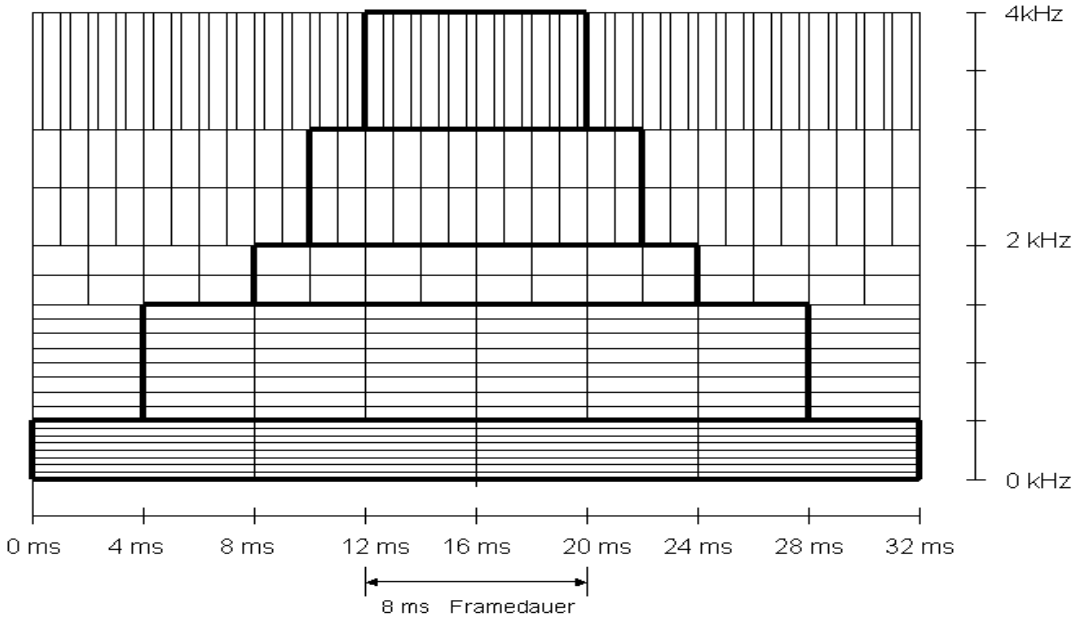

Abbildung 4.3.2: ***Integrationsschema der Teilbandsignale in der Zeit- Frequenzebene. Zur Konstruktion des Merkmalvektors, werden die Abtastwerte innerhalb des markierten Bereichs herangezogen. Auch mit diesem Schema lassen sich die Kurzzeit-Effektivwerte der einzelnen Teilbänder nach der Parseval'schen Beziehung berechnen.***

Bei einer Abtastfrequenz von 8 kHz erhält man bei einer 6-stufigen Zerlegung *L=22* Teilbänder. Die Merkmalvektoren werden bei einer Fensterlänge von $T*N_{l,max}$ = 32 ms und einer Framerate von 8 ms wie folgt konstruiert:

$$e(l) = \frac{1}{N_l} \sum_{n=1}^{N_l} |\, s_l(n) \,|\,, \quad l = 1,2,..., L \qquad (4.3.8)$$

wobei N_l der Anzahl der vom Fenster im Teilband *l* erfassten Abtastwerte $s_l(n)$ entspricht.

Die Teilsummen *e(l)* werden anschließend logarithmiert und mit der *DCT* dekorreliert.

$$SC(k) = \sum_{i=1}^{L} \log(e(l)) \cos\left(\frac{k(l-0.5)}{L}\pi\right), \; k = 1,2...12 \qquad (4.3.9)$$

Die Verfahrensweise der Analyse entspricht der in *[Allen-94]* vorgeschlagenen Methode, d.h. die Teilbänder werden weitgehend unabhängig voneinander verarbeitet. Zu jedem Zeitpunkt können mögliche Änderungen des Signals in allen Skalen gleichzeitig beobachtet werden.

Später wurde in (4.3.8) der Betrag des Signals durch den diskreten ***T**eager **E**nergy **O**perator* ersetzt *[Jabloun-99]*.

$$\Psi_d\ [s(n)] = s^2(n) - s(n+1)s(n-1) \qquad (4.3.10)$$

Insbesondere für Fahrzeuggeräusche konnten gegenüber dem ursprünglichen Ansatz bei Experimenten mit einem System zur Einzelworterkennung nochmals Verbesserungen in der Erkennrate beobachtet werden. *[Kryze-99]* griffen diese Idee auf und verwendeten diesen Ansatz bei einem Experiment zur Phonemerkennung. Allerdings war hier die Abtastfrequenz 16 kHz, was eine Modifikation der Filterbank zur Folge hatte. Neben der Verwendung des *TEO* wurden auch der Logarithmus und als alternative Kompressionskennlinie die fünfte Wurzel untersucht. Die besten Ergebnisse hat man im letzten Fall erhalten, insbesondere bei schlechten SNR < 10 dB zeigte dieser Ansatz um bis zu 10 % bessere Erkennraten als der ebenfalls zum Vergleich herangezogene Vollbandansatz (*PLP*).

4.3.3 Das Auditive Waveletmodell

Diese Ergebnisse motivieren zur Anwendung der Wavelet- Transformation im Perdigao-Modell. Dieses Modell erfordert jedoch eine Berechnung des Effektivwerts in den Teilbändern. Da die Waveletzerlegung eine lineare Transformation ist, gilt auch hier die Parsevalsche Beziehung:

$$\int |\, f(t)\,|^2 = \sum_k |\, a_{0,k}\,|^2 + \sum_j \sum_k |\, b_{j,k}\,|^2 \qquad (4.3.11)$$

Den Effektivwert im *l-ten* Teilband erhält man demnach aus den Wavelet-Koeffizienten $b_{j,k}$ bzw. aus den Teilbandsignalen $s_l(n)$ gemäß:

$$y_l^{RMS} = \sqrt{y_l^P} = \sqrt{\frac{1}{N_l} \sum_{Nl_l} s_l^2(n)} \qquad (4.3.12)$$

Mit diesen Betrachtungen ergibt sich zunächst das folgende Blockschaltbild für ein echtes Teilband-Modell mit dem die Merkmale konsequent entlang der Zeitachse berechnet werden. Im Unterschied zum *PAS*-Modell nach F.Perdigao, werden nun für jedes Teilband Fenster unterschiedlicher Länge benötigt, innerhalb derer für jeden Frametakt jeweils ein *RMS*-Wert berechnet wird. Anschließend folgt eine dynamische Kompression durch das Haarzellenmodell.

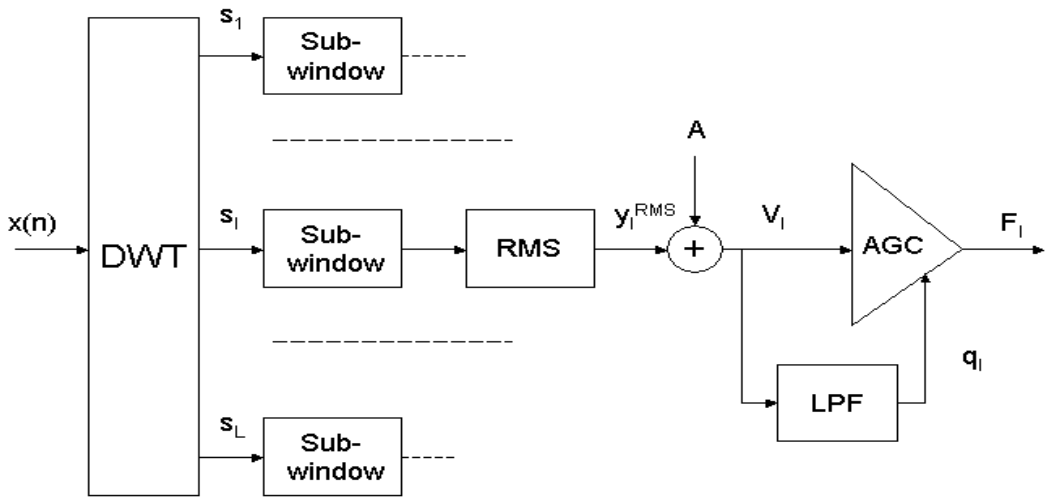

Abbildung 4.3.3 ***Das Auditive Waveletmodell basiert ebenfalls auf dem Perdigao-Modell. Die Kurzzeit-Effektivwerte in den einzelnen Auflösungsstufen werden wiederum nach der Parsevalschen Beziehung berechnet. Darüber hinaus können die Kurzzeit-Effektivwerte auch auf dem TEO basieren.***

In Kapitel 7 wird gezeigt, dass sich für ein vollständiges auditives Modell an diese Stufe noch ein *LIN* anschließt. Das in dieser Arbeit vorgeschlagene *LIN* basiert dabei auf der Berechnung der dominanten Frequenz. D.h. für jeden Kanal wird innerhalb des zugehörigen Teilband-Fensters die dominante Frequenz durch eine Messung der Abstände zwischen den Nulldurchgängen ermittelt. Die Abstände werden ähnlich dem Verfahren von *[Kim-99]* logarithmisch gewichtet und mit der anschließenden Bildung des Kurzeitmittelwerts kann die dominante Frequenz im *l*-ten Teilband zu jedem Frametakt geschätzt werden.

Alternativ wurde von *[Gowdy-00]* vorgeschlagen, die *DCT* der logarithmierten Mel- Filterbank Energie beim *MFCC*- Ansatz durch eine Wavelet-Transformation zu ersetzen. Auch hier kann damit argumentiert werden, dass bei der *DCT* ebenfalls Basisfunktionen verwendet werden, welche das gesamte Mel-Frequenzband abdecken, demnach werden lokale Störungen ebenfalls alle *MFCC* Koeffizienten beeinflussen. Außerdem wird bei der Berechnung der *DCT* implizit ein Rechteckfenster fester Länge verwendet, so dass auch hier eine feste Frequenz- und Zeitauflösung vorliegt. Eine weitere Möglichkeit zur Dekorrelation benachbarter Teilbänder ist die Anwendung eines *LIN*. Ausgehend von den Arbeiten von *[Wang-94]* und *[Ghitza-94]* kann ein *LIN* konstruiert werden, welches in Abhängigkeit des Abstandes von der dominanten Frequenz im aktuellen Kanal zu dessen Mittenfrequenz einen Gewichtsfaktor berechnet. Mit diesem Faktor kann die Abhängigkeit benachbarter Kanäle reduziert werden.

5 Auditive Modelle mit CAS Eigenschaften

Im Folgenden beziehen sich alle Ausführungen auf das Modell im Frequenzbereich, da bisher nur für dieses Modell eine relativ einfache recheneffiziente Implementierung gefunden wurde. Es werden zwei Ansätze verfolgt. Der erste Ansatz bewertet die kohärenten Signalkomponenten erst in der Teilbandebene nach dem *LIN*, im zweiten Ansatz dagegen erfolgt die Bewertung kohärenter Signalkomponenten bereits in der Frequenzebene. In beiden Fällen erfolgt die Bestimmung zueinander kohärenter Signalkomponenten unter Verwendung der Sprachgrundfrequenz und deren Harmonischen.

Weitere Verbesserungen der Zuverlässigkeit können dann erwartet werden, wenn auch wesentliche Eigenschaften des Zentralen Auditiven Systems (*engl.* ***C****entral* ***A****uditory* ***S****ystem*) in die Modellierung einfließen. Eine dieser Eigenschaften ist die so genannte Bindungseigenschaft. Unter diesem Begriff wird die Fähigkeit verstanden, zusammengehörige Signalkomponenten eines auditiven Objekts hervorzuheben und vom Hintergrund zu trennen. Eine wichtige Voraussetzung zur Bildung von auditiven Objekten besteht zunächst darin, zusammengehörige Signalkomponenten identifizieren zu können. In der einschlägigen Fachliteratur sind hierzu einige Vorschläge beschrieben worden, so könnte man sich bspw. auf diejenigen Signalkomponenten konzentrieren, welche gleichzeitig einsetzen und sich anschließend relativ langsam und kohärent entwickeln (z.B. die Harmonischen der Sprachgrundfrequenz). Ein solcher Ansatz wurde in *[Andringa-02]* beschrieben, dieser ist allerdings mit einem hohen Rechenaufwand verbunden und derzeit nicht in einem echtzeitfähigen *ASR*-System umsetzbar.

In einem ähnlichen Ansatz wird als möglicher Indikator für die Zusammengehörigkeit von Signalkomponenten die gemeinsame Einhüllende von Teilbandsignalen betrachtet *[Hasegawa-97]*. Eine gemeinsame Einhüllende der Teilbandsignale kann in den stimmhaften Abschnitten gefunden werden. Deren Ursprung ist wiederum auf die Sprachgrundfrequenz zurückzuführen. Werden nun diejenigen Signalkomponenten, welche sich durch eine gemeinsame Einhüllende auszeichnen besonders betont, dann führt dies auf so genannte pitchkohärente Merkmale (*engl.* ***P****itch* ***C****oherent* ***F****eatures*).

5.1 Modellierung der Bindungseigenschaft im CAS Modell

5.1.1 Bindungseigenschaft

Der stimmhafte Anteil von Sprache wird durch eine Menge von periodischen Komponenten (Formanten) repräsentiert, deren Einhüllende wird jeweils durch die Sprachgrundfrequenz (Pitch) bestimmt. Andererseits ist bekannt, dass synchron feuernde Spikes verschiedene Merkmale eines einzigen wahrgenommenen Objekts codieren. Dieses Phänomen wird als Bindungseigenschaft bezeichnet. Nun lässt sich daraus die Hypothese ableiten, dass dieses Prinzip auch auf höheren Ebenen zur Anwendung kommt. Die Bindung von zusammengehörigen Signalkomponenten könnte demnach von der Sprachgrundfrequenz geleistet werden. Im Folgenden wird daher ein Verfahren vorgestellt, mit dem die stimmhaften Komponenten über die Sprachgrundfrequenz identifiziert werden und anschließend gegenüber dem Geräuschhintergrund angehoben werden.

5.1.2 Primitive Bindung

Die Berechnung pitchkohärenter Merkmale basiert auf dem Auditiven *PAS*-Modell im Frequenzbereich. Mit den Teilbandspektren $|H_i(l)|^2|X(l)|^2$ können unter Verwendung der *IFFT* die Teilband-AKF berechnet werden. Dabei gelten folgende Zusammenhänge *(l-Frequenzindex, m-Verschiebungsindex*):

$$R[m] \xrightarrow{Fourier} |X(l)|^2 \sum_{i=1}^{N} |H_i(l)|^2 \quad \text{,mit} \quad R[m] = \sum_{i=1}^{N} r_i[m] \tag{5.1.1}$$

Ausgehend von der Gesamtband-AKF *R[m]* kann zunächst die Periode der Sprachgrundfrequenz geschätzt werden:

$$M = \arg \max_{m} \{R[m]\} \tag{5.1.2}$$

Bei einer Abtastfrequenz von 8 kHz entspricht der für die Perioden M zulässige Bereich einem Sprachgrundfrequenzbereich von 62,5...400 Hz. Anschließend erfolgt für jedes Teilband die Berechnung eines normierten AKF-Werts an der Stelle M (Voice-Index).

$$g_i^M(k) = \frac{r_i[M]}{r_i[0]} \qquad i: Teilbandindex;\ \ k: Frameindex \tag{5.1.3}$$

Alle derart ermittelten Gewichtungsfaktoren werden zu einem Vektor $\mathbf{g}^M(k)$ zusammengefasst, mit dem nun eine pitchkohärente Gewichtung des *PAS*-Spektrums erreicht werden kann. Darüber hinaus kann unter Verwendung der Gleichungen (5.1.1) und (5.1.2) für jeden Frame ein Maß für dessen Stimmhaftigkeit *(engl. Voicedness)* angegeben werden.

$$v^M(k) = \frac{R[M]}{R[0]} \qquad k: Frameindex \tag{5.1.4}$$

Damit ergibt sich das vorläufige *CAS* Spektrum zunächst durch Multiplikation der Gewichtungsfaktoren mit den zugehörigen Komponenten des *PAS*-Spektrum:

$$CAS'(k) = \mathbf{g}^M(k).*PAS(k) \tag{5.1.5}$$

Auch in stimmlosen Abschnitten erfolgt hier eine Bindung der Teilbänder. Dies hat zur Folge, dass dann die Gewichtungsvektoren als multiplikative Rauschsignale wirksam werden. Deshalb wird diese Umsetzung der Bindungseigenschaft als primitive Bindung bezeichnet.

5.1.3 Bedingte Bindung

Unter Verwendung der normierten Stimmhaftigkeit (*engl.* Voicedness) kann eine bedingte Bindung der Teilbänder durch eine einfache Linearkombination erfolgen.

$$CAS'(k) = v(k)*[\mathbf{g}^M(k).*PAS(k)] + [1-v(k)]*PAS(k) \tag{5.1.6}$$

Um verbleibende Reststörungen unterdrücken zu können, wird das *CAS*-Spektrum abschließend einem Amplitudenmodulationstiefpass zugeführt. Dieser Ansatz kann damit erklärt werden, dass ein Kurzzeit-*RMS*-Mittelwert grundsätzlich auch als Einhüllendendetektor aufgefasst werden kann. Die Grenzfrequenz des AM-Tiefpass wird wie bei *[Tchorz-99]* oder *[Hermansky-94]* auf etwa 8 Hz festgelegt. Damit werden nur noch Sprachanteile betrachtet, die sich mit der Silbenfrequenz entwickeln.

Da alle Berechnungen nach der *FFT* mit Framerate ausgeführt werden, kann die *AM*-Filterung in Zustandsform realisiert werden:

$$CAS(k) = f_{AM}[Z(k), CAS'(k)] \qquad (5.1.7)$$

Diese Darstellungsform wird später noch etwas ausführlicher erläutert, zunächst soll es genügen, dass sich das *CAS*-Spektrum aus den aktuellen Zuständen der *AM*-Filter und dem aktuellen Eingangsspektrum *CAS'(k)* ergibt.

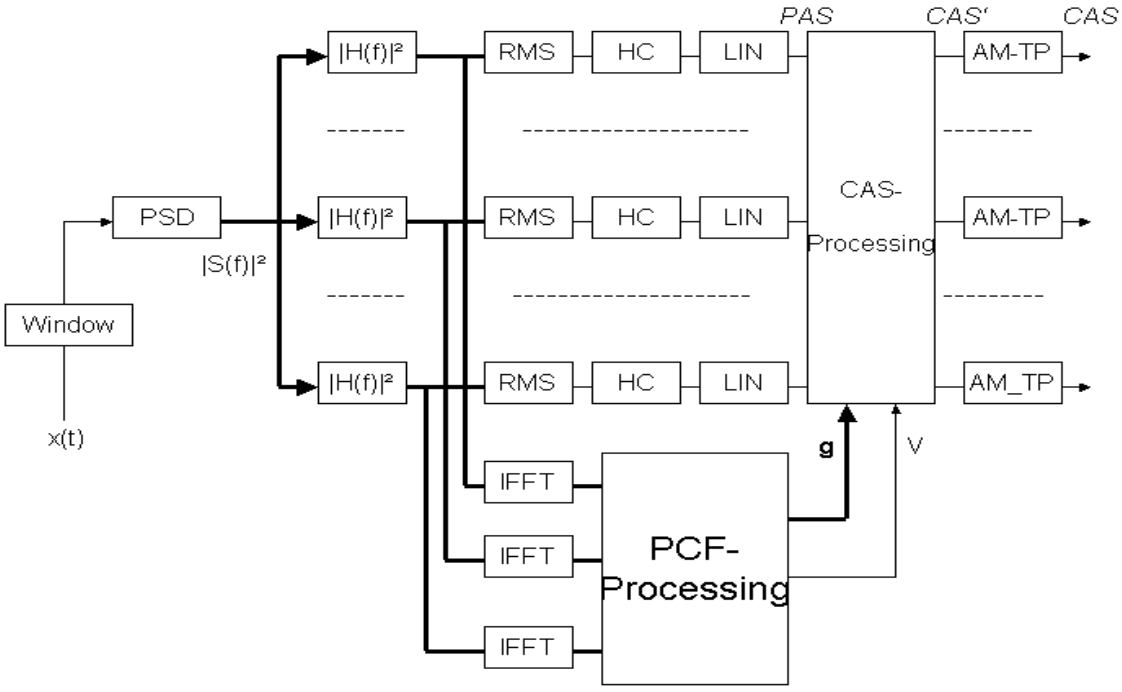

Abbildung 4.3.3 ***Kombiniertes PAS-CAS-Modell. Der obere Zweig der Abbildung entspricht bis einschließlich der Lateralen Inhibition dem PAS-Modell. Im unteren Zweig werden für jeden Frame die Stimmhaftigkeit und der pitchkohärente Gewichtungsvektor berechnet. Im Block CAS-Processing erfolgt dann die Berechnung des vorläufigen CAS-Spektrums, verbleibende Reststörungen größer als 8 Hz werden vom AM-Tiefpass unter-drückt.***

5.2 Einbindung der Vorhersage zur Auflösung von Mehrdeutigkeiten

Motivation

Die Zuverlässigkeit des gebundenen auditiven Spektrums wird maßgeblich von der Schätzqualität der zur Sprachgrundfrequenz F_0 gehörenden Periode *M* bestimmt. Akustische Störungen oder Hintergrundsprecher können zu Mehrdeutigkeiten oder zu sprunghaften Änderungen von F_0 führen und damit die Schätzqualität beeinträchtigen.

Nun ist aber bekannt, dass der Informationsfluss zwischen den Komponenten Sinneszellen, *PAS* und *CAS* sowohl auf afferenten als auch auf efferenten Nervenbahnen stattfindet. Es ist denkbar, dass auf den efferenten Nervenbahnen Vorhersagen zurückgeführt werden, um Mehrdeutigkeiten aufzulösen oder Sprünge zu unterbinden.

Prinzip

Durch eine Vorhersage des zu erwartenden Spektrums innerhalb stimmhafter Abschnitte wird nun versucht, sich kontinuierlich entwickelnde *CAS*-Spektren zu erzwingen. Zunächst wird eine geordnete Liste von *L* möglichen Pitchperioden erzeugt. Nach Vorliegen eines aktuellen *PAS*-Spektrums erfolgt dessen Variation unter Verwendung der *L*-pitchkohärenten Gewichtungsvektoren. Auf der Grundlage einer Prädiktion wird eines der möglichen Spektren ausgewählt und dem *AM*-Tiefpass zugeführt. Für den nächsten Zeitschritt wird das *CAS*-Spektrum (Ausgänge der *AM*-Tiefpässe) auf den Prädiktor zurückgeführt. Für die Vorhersage kann ein Kalmanprädiktor verwendet werden, wenn der *AM*-Tiefpass als Minimalphasensystem in Zustandsdarstellung realisiert wird.

Ein Nachteil des unter 5.1 beschriebenen Verfahrens besteht darin, dass selbst in stimmhaften Abschnitten und (oder) bei akustischen Störungen die Periode *M* sprunghaften Änderungen unterliegen kann. Eine Lösung dieser Problematik kann durch die folgende Beobachtung am biologischen Vorbild gewonnen werden:

Wie bereits in der Einführung erwähnt, haben Untersuchungen an Tieren gezeigt, dass sich insbesondere die Erkennung von vokalischen Lauten deutlich verschlechtert, wenn die efferenten Nervenbahnen unterbrochen werden *[Dusan-05]*. Zudem ist aus der Hirnforschung bekannt, dass ein inneres Abbild der motorischen Strukturen und deren Dynamik ausgenutzt werden kann, um Eingangssignale vorherzusagen. Nun kann man vermuten, dass auf den efferenten Bahnen Vorhersagen bezüglich des zu erwartenden Eingangssignals zurückgeführt werden und dass diese Vorhersagen in stimmhaften Abschnitten stärker berücksichtigt werden als in stimmlosen Abschnitten. Daher erfolgt auch für das *PAS-CAS-Modell* mit Kalmanvorhersage eine weiche Unterscheidung zwischen stimmhaften und stimmlosen Abschnitten.

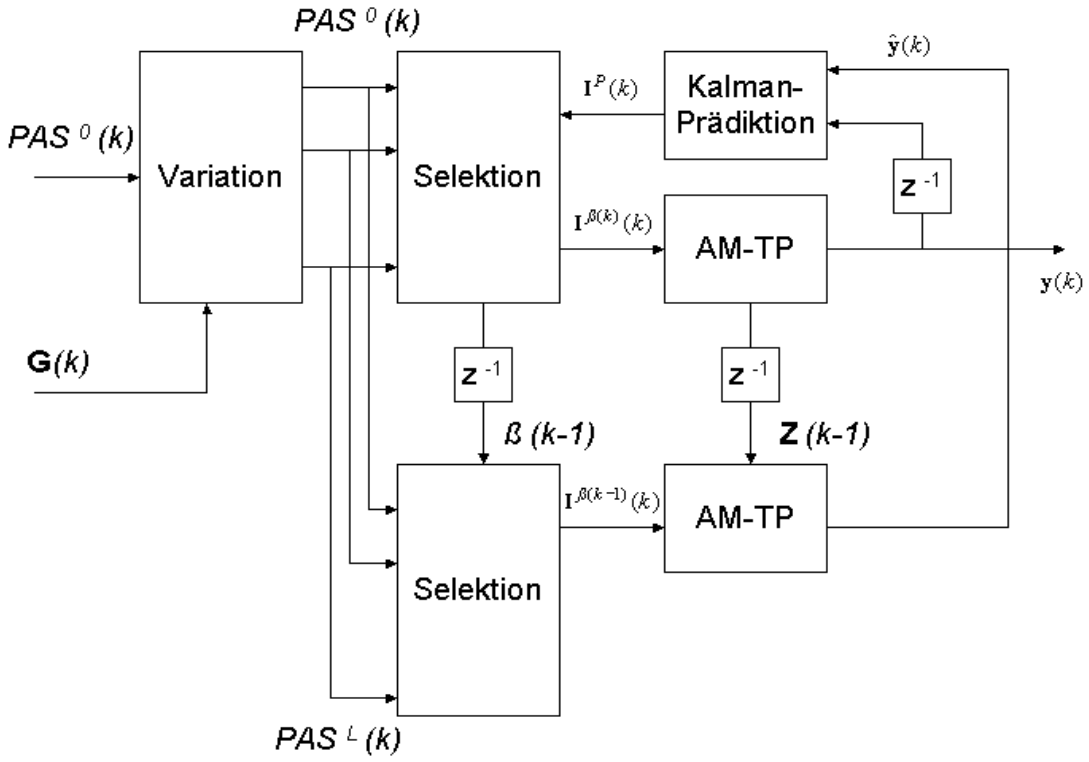

Abbildung 5.2 ***Kombiniertes PAS-CAS-Modell mit Kalmanprädiktion: Ausgehend von den eingehenden Gewichtungs-vektoren und dem PAS-Spektrum werden L-Variationen erzeugt. Unter Verwendung der Kalman-vorhersage wird nun diejenige Variation ausgewählt, welche am stärksten der Kontinuitätsbedingung gehorcht.***

Unter Verwendung von Abbildung 5.2 kann nun die Struktur des *PAS-CAS*-Modells mit Kalmanvorhersage näher erläutert werden: Nach Vorliegen eines aktuellen *PAS*-Spektrums erfolgt dessen Variation unter Verwendung einer festen Anzahl von pitchkohärenten Gewichtungsvektoren. Nach Auswahl einer *PAS*-Variation auf der Grundlage der Kalmanvorhersage und einer anschließenden AM-Tiefpassfilterung werden die Ausgänge aller AM-Tiefpässe wiederum dem Kalmanprädiktor zugeführt. Dieser kann dann eine Vorhersage des Eingangsspektrums für den nächsten Zeitschritt liefern. Im Kern ähnelt dieses Verfahren einem ***E**volutionär **O**rientierten **A**lgorithmus*, der sich aus den Komponenten Variation, Selektion und Reproduktion zusammensetzt. Aus informationstheoretischer Sicht liegen die Vorzüge des *EOA* vor allem in der zuverlässigen Reproduktion oder der Erhaltung von akustischen Eigenschaften eines auditiven Objekts bei gleichzeitiger Überlagerung von akustischen Störungen oder anderen auditiven Objekten.

5.2.1 Variation des PAS Spektrums

Um sprunghafte Änderungen im Systemzustand zu vermeiden, berücksichtigt das nun vorgeschlagene Verfahren nicht nur das globale Maximum und dessen zugehörige Pitchperiode M sondern eine geordnete Liste von lokalen Maxima.

$$\mathbf{M}(k) = [M^1(k), M^2(k), ..., M^L(k)] \;, \qquad L : Anzahl\ der\ Variationen \tag{5.2.1}$$

Wie in Abschnitt 5.1 beschrieben wurde, ist jedem *PAS*-Teilband ein auf den Bereich von 0-8 Hz beschränkter Amplitudenmodulationstiefpass nachgeschaltet. Wird der entsprechende Modulationstiefpass als *IIR*-Filter in der Direktform *II* realisiert, so bezeichnet man den Signalvektor in der Verzögerungskette des Filters als den aktuellen inneren Zustand. In ihrer Gesamtheit könnte man dann alle inneren Zustände der Modulationstiefpässe als ein verborgenes Abbild des Zustands des Artikulationsapparates interpretieren. Da sich der innere Zustand des Artikulationsapparates nicht sprunghaft ändern sollte, kann ein glatter Verlauf der Zustandstrajektorie wie folgt erzwungen werden: Aus (5.2.1) folgt, dass für jeden Frame *L*- Gewichtungsfaktoren $\mathbf{G}(k) = [\mathbf{g}^1(k), \mathbf{g}^2(k), ..., \mathbf{g}^L(k)]$ gemäß (5.1.3) und *L*-Stimmhaftigkeiten $\mathbf{V}(k) = [v^1(k), v^2(k), ..., v^L(k)]$ gemäß (5.1.4) berechnet werden können. Unter Verwendung der *L*-Gewichtungsvektoren folgt nun eine Variation des *PAS*-Spektrums. Zusätzlich wird ein Vektor $\mathbf{g}^0(k)$ eingeführt, dessen Komponenten alle den Wert „1“ erhalten.

$$PAS^l(k) = PAS(k) .* \mathbf{g}^l(k) \;\; ; \;\; l = (0...L) \tag{5.2.2}$$

Aus den Variationen (*1…L*) kann bei Vorlage der Kalmanvorhersage in den stimmhaften Abschnitten dasjenige Spektrum gewählt werden, welches am stärksten der Kontinuitätsbedingung gehorcht. In den stimmlosen Abschnitten dagegen wird den *AM*-Tiefpässen das unveränderte Spektrum $PAS^0(k)$ zugeführt.

5.2.2 Selektion

Mit der Kalmanvorhersage $\mathbf{I}^P(k)$ und den Variationen $PAS^l(k)$ kann nun gemäß der „Kontinuitätsbedingung“ diejenige *PAS*-Variation ausgewählt werden, welche den geringsten mittleren quadratischen Fehler bezüglich der Vorhersage aufweißt. Der mittlere quadratische Fehler kann als Kosten- oder „Fitness“- Funktion bezeichnet werden. Formal erfolgt die Selektion nach der Vorschrift:

$$ß(k) = \underset{l=1..L}{\arg\min} \left[\| PAS^l(k) - \mathbf{I}^P(k) \|^2 \right] \tag{5.2.3}$$

Das selektierte Spektrum zum Zeitschritt k, d.h. das aktuelle Eingangsspektrum der *AM*-Tiefpässe wird mit $PAS^{ß(k)}(k)$ bezeichnet.

Weiter oben wurde bereits angesprochen, dass die Vorhersagen lediglich auf die stimmhaften Abschnitte beschränkt bleiben sollten. Um Diskontinuitäten in den Zustandsentscheidungen (stimmhaft, stimmlos) zu vermeiden, wird deshalb auf eine weiche Entscheidung zurückgegriffen:

$$\mathbf{I}^{ß(k)}(k) = v^{\beta(k)}(k)\, PAS^{ß(k)} + (1 - v^{ß(k)}(k))\, PAS^{0}(k) \qquad (5.2.4)$$

Das *AM*-Eingangsspektrum $\mathbf{I}^{ß(k)}(k)$ ergibt sich demnach wiederum als gewichtete Summe des stimmhaften Spektrums und des ursprünglichen Spektrums.

5.2.3 Reproduktion

Um eine Vorhersage des Eingangssignals auf der Basis eines Kalmanfilters ausführen zu können, muss das Amplitudenmodulationsfilter als Minimalphasenfilter in Zustandsform implementiert werden. Daher erfolgt in diesem Abschnitt zunächst eine Einführung in das Konzept der Zustandsvariablen. Anschließend wird die Realisierung des Reproduktionsschritts beschrieben. Die Beschreibung eines diskreten Systems basiert häufig auf einer Differenzengleichung *N*-ter Ordnung:

$$x(k)b'_0 + x(k-1)b'_1 + ... + x(k-N)b'_N = y(k) + y(k-1)a'_1 + ... + y(k-N)a'_N \qquad (5.2.5)$$

Die Übertragungsfunktionen für dieses System und dem zugehörigen inversen System können in der folgenden Form angegeben werden:

$$H(z) = \frac{\sum_{k=0}^{N} b'_k\, z^{-k}}{1 + \sum_{k=1}^{N} a'_k\, z^{-k}} \quad ; \quad H^{INV}(z) = \frac{\frac{1}{b'_0}\sum_{k=0}^{N} a'_k\, z^{-k}}{1 + \sum_{k=1}^{N} \frac{b'_k}{b'_0} z^{-k}} = \frac{\sum_{k=0}^{N} a_k z^{-k}}{1 + \sum_{k=1}^{N} b_k z^{-k}} \qquad (5.2.6)$$

Liegen sowohl die Polstellen als auch die Nullstellen von *H(z)* innerhalb des Einheitskreises, so erhält man ein stabiles Minimalphasensystem. Die Einführung von Zustandsvariablen lässt sich darauf zurückführen, dass jede Differenzengleichung *N*-ten Grades in eine Vektordifferenzengleichung ersten Grades überführt werden kann. Für diskrete Systeme die in der Direktform 2 dargestellt werden, ist die Interpretation der Zustandsvariablen besonders einfach. Hier werden die Signalwerte in den Verzögerungsgliedern als Zustandsvariablen aufgefasst und in einem Zustandsvektor ***z(k)*** zusammengefasst. Für ein lineares zeitinvariantes diskretes System lauten die Zustandsgleichungen für das System *H(z) [Pollock-99]:*

$$\mathbf{z}'(k) = \mathbf{A}\mathbf{z}'(k-1) + \mathbf{C}x(k) \qquad (5.2.7)$$

$$y(k) = \mathbf{B}\mathbf{z}'(k)$$

Die erste Gleichung beschreibt die Berechnung des aktuellen Zustands, die zweite Gleichung wird als Messgleichung bezeichnet. Zur Berechnung des aktuellen Systemausgangs ist demnach nur der alte Zustandsvektor und der aktuelle Eingangswert notwendig. Die Zustandsgleichungen für das inverse System lauten:

$$\mathbf{z}(k) = \mathbf{B}\mathbf{z}(k-1) + \mathbf{C}y(k) \qquad (5.2.8)$$

$$x(k) = \mathbf{A}\mathbf{z}(k)$$

Die Koeffizienten des *IIR*-Systems in Gleichung (5.2.5) werden so gewählt, dass die Bedingungen Tiefpasscharakteristik und Minimalphasensystem eingehalten werden. Ein Reproduktionsschritt besteht nun darin, dass in Gleichung (5.2.7) der ausgewählte Vektor $\mathbf{I}^{ß(k)}(k)$ für *x(k)* eingesetzt wird und zusammen mit dem alten Zustandsvektor der jeweilige Ausgangswert *y(k)* berechnet wird. Die Ausgänge aller *AM*-Tiefpässe zum Zeitpunkt *k* werden zu einem Vektor $\mathbf{y}(k)$ zusammengefasst, darüber hinaus wird der alte Zustand des *AM*-Tiefpass später für die Kalmanvorhersage benötigt.

5.2.4 Kalmanvorhersage

Mit dem Einsatz eines Kalmanalgorithmus besteht nun die Möglichkeit das zu erwartende PAS- Spektrum vorherzusagen. Zu diesem Zweck werden gemäß Abb. 5.2 die Ausgänge aller Reproduktionsfilter den Kalmanprädiktoren in den Rückwärtszweigen zugeführt. Das für die Kalmanvorhersage benötigte Modell basiert auf den Zustandsgleichungen des inversen AM-Tiefpasses. Zunächst folgt die Beschreibung des Kalman-Standardverfahrens für das inverse Filter *[Pollock-99]*, anschließend werden die Gleichungen für eine Mehrschritt-Prädiktion angegeben. Die Teilbandindizes wurden der Einfachheit halber weggelassen.

$$\begin{aligned}
\bar{\mathbf{z}}(k) &= \mathbf{B}\,\hat{\mathbf{z}}(k-1) + \mathbf{C}\,y(k) \\
\mathbf{P}^{-}(k) &= \mathbf{B}\mathbf{P}(k-1)\mathbf{B}^{T} + Q_1 \\
\tilde{x}(k) &= \mathbf{A}\,\bar{\mathbf{z}}(k) \\
\mathbf{G}(k) &= \mathbf{P}^{-}(k)\mathbf{A}^{T}[\mathbf{A}\mathbf{P}^{-}(k)\mathbf{A}^{T} + Q_2]^{-1} \\
\hat{\mathbf{z}}(k) &= \bar{\mathbf{z}}(k) + \mathbf{G}(k)[x^{ß}(k) - \tilde{x}(k)] \\
\mathbf{P}(k) &= [\mathbf{I} - \mathbf{G}(k)\mathbf{A}]\mathbf{P}^{-}(k)
\end{aligned} \qquad (5.2.9)$$

In Abb. 5.2 wurde berücksichtigt, dass verzögerungsfreie Rückführungen nicht zulässig sind. Das Kalmanfilter erhält also im Zeitschritt *k* als Eingangsgröße den Wert *y(k-1)* und erzeugt demnach den Vorhersagewert $\tilde{x}(k-1)$, gesucht ist allerdings der Vorhersagewert $\tilde{x}(k)$. Dieser Wert kann durch Mehrschritt-Prädiktion ebenfalls gemäß *[Pollock-99]* berechnet werden. Dabei wird zusätzlich der zukünftige Zustand und anschließend der Vorhersagewert berechnet.

$$\begin{aligned}
\hat{\mathbf{z}}(k) &= \mathbf{B}^{2}\hat{\mathbf{z}}(k-1) + \mathbf{B}\mathbf{C}y(k-1) + \mathbf{C}y(k) \\
\tilde{x}(k) &= \mathbf{A}\,\hat{\mathbf{z}}(k)
\end{aligned} \qquad (5.2.10)$$

Eine Betrachtung von Abbildung 5.2 zeigt, dass der Wert *y(k)* noch nicht vorliegt, er kann erst nach der Selektion berechnet werden. Ein Blick auf den unteren Signalzweig zeigt aber, dass $\hat{y}_i(k)$ - die Antwort auf die Komponente *i* der Variation $\mathbf{I}^{ß(k-1)}(k)$ - als Schätzung für $y_i(k)$ verwendet werden kann. Die Vorhersagen $\tilde{x}(k)$ aller Teilbänder werden abschließend in einem Vektor $\mathbf{I}^{P}(k)$ zusammengefasst. Dieser Vektor liefert die Vorhersage für das zu erwartende Eingangsspektrum und unterstützt eine zuverlässigere Reproduktion bzw. die Erhaltung von akustischen Eigenschaften des identifizierten auditiven Objekts entlang der Zeitachse.

5.3 Hochauflösende CAS Modellierung

Bei dieser Modellierung wird die Bindungsoperation von der grob auflösenden *RMS*-Teilbandebene in den hoch auflösenden Frequenzbereich verlagert. Zur Schätzung der Sprachgrundfrequenz kann nun einerseits auf das geräuschreduzierte Spektrum zurück gegriffen werden, andererseits wird nur noch eine einzige *IFFT* benötigt wird. Die Spektrale Subtraktion muß hier im Frequenzbereich durchgeführt werden. Um eine genügende Auflösung für die Sprachgrundfrequenz zu erhalten, muß man die Anzahl der *FFT* Stützstellen von 256 auf 1024 erhöhen. Die Bindungsoperation kann hier wieder als primitive oder bedingte Bindung ausgeführt werden. Die Bestimmung der Sprachgrundfrequenz erfolgt also ähnlich wie in Abschnitt 5.1 über die *AKF* des geräuschreduzierten Kurzzeit-Leistungsspektrums.

$$|S(k)|^2 \overset{IFFT}{\rightarrow} R[m] \tag{5.3.1}$$

Mit dem Kehrwert der gefundenen Periodendauer *M* kann die Übertragungsfunktion des adaptiven Kammfilters nach einem Ansatz von *[Vondra-07]* berechnet werden.

$$H(f) = \max\left\{\frac{\exp[x\cos 2\pi f / F_0)]}{y}, 10^{\mu/20}\right\} \tag{5.3.2}$$

wobei $H_{\max} = e^x / y$ und $H_{\min} = 10^{\mu/20}$ *für* $0 \le f \le F_S/2$ gilt. Die Parameter *x, y* und μ sind Konstanten, sie bestimmen die Bandbreite und das Betragsmaximum für jede Keule des Kammfilters. Der Parameter μ ergibt sich bei Vorgabe des Betragsminimums zu $\mu = 20\log H_{\min}$. Hier erhält man die Variationen durch Multiplikation des geräuschreduzierten Spektrums mit den Kammfilterfunktionen der verschiedenen F_0 Kandidaten. Mit dem Vorliegen der *AKF* kann die Stimmhaftigkeit *v* wieder nach Gleichung (5.1.4) berechnet werden. Die im Abschnitt 5.2 beschriebenen Bindungstypen, können daher auch im Frequenzbereich realisiert werden.

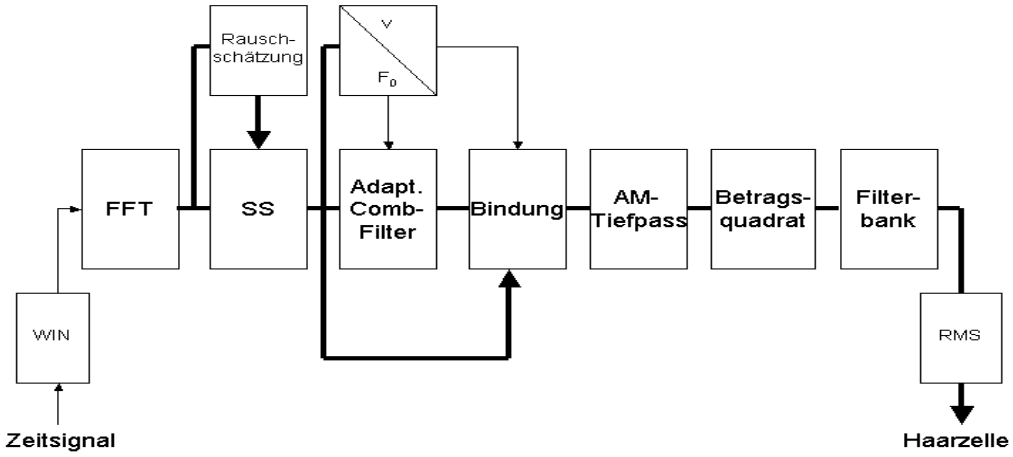

Abbildung 5.3 ***Hochauflösendes PAS-CAS-Modell: Sowohl die Gewichtungsfunktion, als auch die Bindungsoperation liegt hier noch vor der Filterbank. Zur Schätzung der Sprachgrundfrequenz wird die Anzahl der FFT Stützstellen von 256 auf 1024 erhöht (Abtastfrequenz 8 kHz).***

5.4 Hochauflösende CAS Modellierung mit Vorhersage

Ähnlich wie in Abschnitt 5.2 soll auch hier das Kontinuitätsprinzip eingehalten werden. Dabei kommt wiederum der *Evolutionär Orientierte Algorithmus* zur Anwendung. Da die Prädiktion nun auf der Frequenzebene stattfindet (512 Stützstellen) wird hier aus Rechenzeitgründen auf eine Kalmanprädiktion auf der Grundlage des inversen *AM*-Tiefpass verzichtet und eine einfachere Methode vorgeschlagen. Eine grobe Schätzung des vorherzusagenden Ausgangswerts kann man aus den Zustandsgleichungen des *AM*-Tiefpass ableiten.

$$\begin{aligned} \mathbf{z}(k+1) &= \mathbf{A}\mathbf{z}(k) + \mathbf{C}x(k) \\ y(k) &= \mathbf{B}\mathbf{z}(k) \\ \hat{y}(k+1) &= f(\mathbf{z}(k+1), x(k)) \end{aligned} \qquad (5.4.1)$$

Die Schätzwerte aller Frequenzstützstellen können nun dem Selektionsblock in Abbildung 5.4 zur Verfügung gestellt werden. Das aus den Variationen auszuwählende Spektrum hat dann den geringsten Abstand zur Prädiktion und gehorcht somit dem Kontinuitätsprinzip. Schließlich wird die Linearkombination aus geräuschreduzierten Spektrum und pitchkohärentem Spektrum dem AM-Tiefpass zur Reproduktion übergeben. Die nachfolgenden Blöcke sorgen dafür, dass der Haarzelle *RMS*-Werte zugeführt werden.

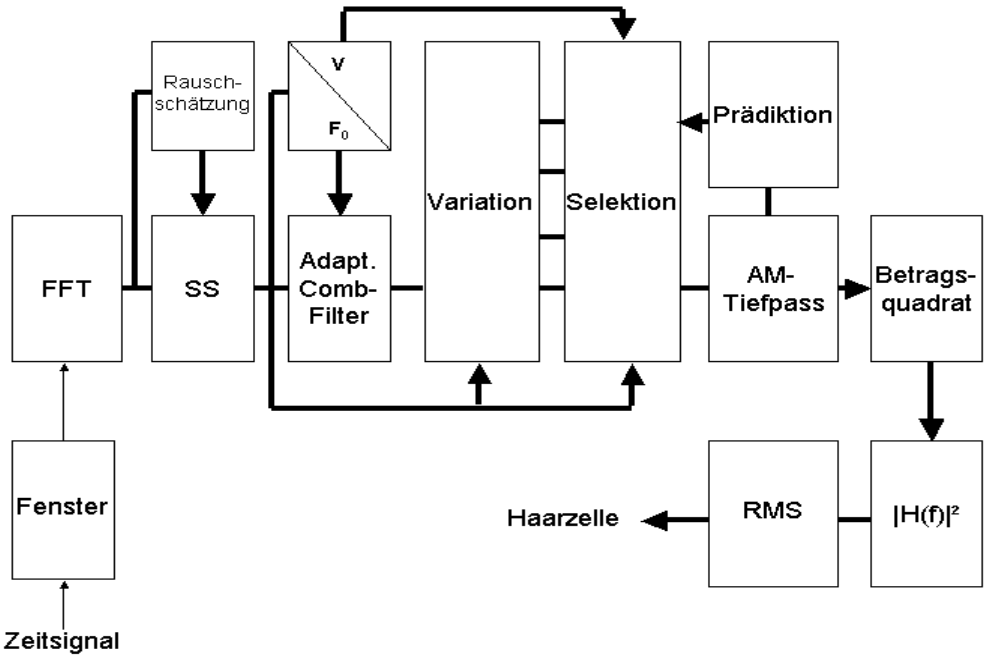

Abbildung 5.4 ***Hochauflösendes PAS-CAS-Modell mit Vorhersage: Die Bindung erfolgt hier gemäß dem Prinzip Variation, Selektion und Reproduktion. Zur Selektion werden Vorhersagen benötigt, diese können unter Verwendung der Zustandsraumdarstellung vom AM- Tiefpass abgegriffen werden. Auch bei diesem Modell wurde die Anzahl der FFT Stützstellen von 256 auf 1024 erhöht (Abtastfrequenz 8 kHz).***

6 Verfahren zur Geräuschunterdrückung

6.1 Allgemeine Betrachtungen

In diesem Kapitel werden verschiedene Rauschunterdrückungsmethoden für auditive Modelle vorgestellt, auf die Notwendigkeit einer genügend genauen Rauschschätzung wird ebenfalls eingegangen. Einführend werden einige grundlegende Zusammenhänge erläutert.

Gemäß einer linearen Modellvorstellung durchläuft das Sprachsignal einen Übertragungskanal, welcher eine Änderung des Amplituden- und Phasenspektrums des zu übertragenden Sprachsignals zur Folge hat. Im Frequenzbereich führt das zu der Darstellung:

$$X(f) = H(f)S(f) = |H(f)S(f)|\, e^{j[\angle H(f) + \angle S(f)]} \tag{6.1.1}$$

Ein weiterer Bestandteil der Modellvorstellung ist die Überlagerung des Nutzsignals mit einem Störsignal *N(f).*

$$Y(f) = H(f)S(f) + N(f) \tag{6.1.2}$$

Mathematisch kann dieser Prozess als eine affine lineare Transformation verstanden werden, welche in Matrixschreibweise durch Gleichung (6.1.3) angegeben werden kann:

$$\mathbf{y} = \mathbf{H}\mathbf{s} + \mathbf{n} \tag{6.1.3}$$

Das Sprachsignal erhält man im Prinzip durch die folgenden Operationen wieder zurück:

$$\mathbf{s} = \mathbf{H}^{-1}(\mathbf{y} - \mathbf{n}) \tag{6.1.4}$$

Dieser Ansatz fordert zum einen die Schätzung der Störung ***n*** und zum anderen die Schätzung der inversen Kanalmatrix $\boldsymbol{H}^{-1}$. Nun ist es in der Sprachverarbeitung zwar üblich eine Subtraktion des Störsignals im Frequenzbereich vorzunehmen, die Reduktion der Kanaleinflüsse geschieht aber häufig durch eine cepstrale Subtraktion.

Die nachfolgenden Betrachtungen konzentrieren sich jedoch nur auf die Überlagerung von Störgeräuschen, dabei kann man sich die Störung *N(f)* aus den folgenden zwei Komponenten zusammengesetzt denken:

$$N(f) = E\{N(f)\} + \Delta N(f) \tag{6.1.5}$$

In der Praxis beschränkt man sich fast immer auf die Betragsdarstellung des komplexen Spektrums oder auf dessen Betragsquadrat. Um von diesen beiden Varianten zu abstrahieren, wird im Folgenden die komplexe Schreibweise beibehalten.

Die am häufigsten verwendete Methode zur Rauschminderung - die Spektrale Subtraktion - beschränkt sich zunächst nur darauf, den Erwartungswert der Störung zu schätzen und vom empfangenen Spektrum zu subtrahieren.

$$X'(f) = Y(f) - E\{N(f)\} \tag{6.1.6}$$

Darüber hinaus kann aber auch die Standardabweichung der Reststörung geschätzt und subtrahiert werden.

$$X''(f) = X'(f) - STD\{\Delta N(f)\} \tag{6.1.7}$$

Abbildung 6.1 fasst noch einmal alle bisher genannten Operationen und die dazugehörigen Auswirkungen im Signalraum zusammen.

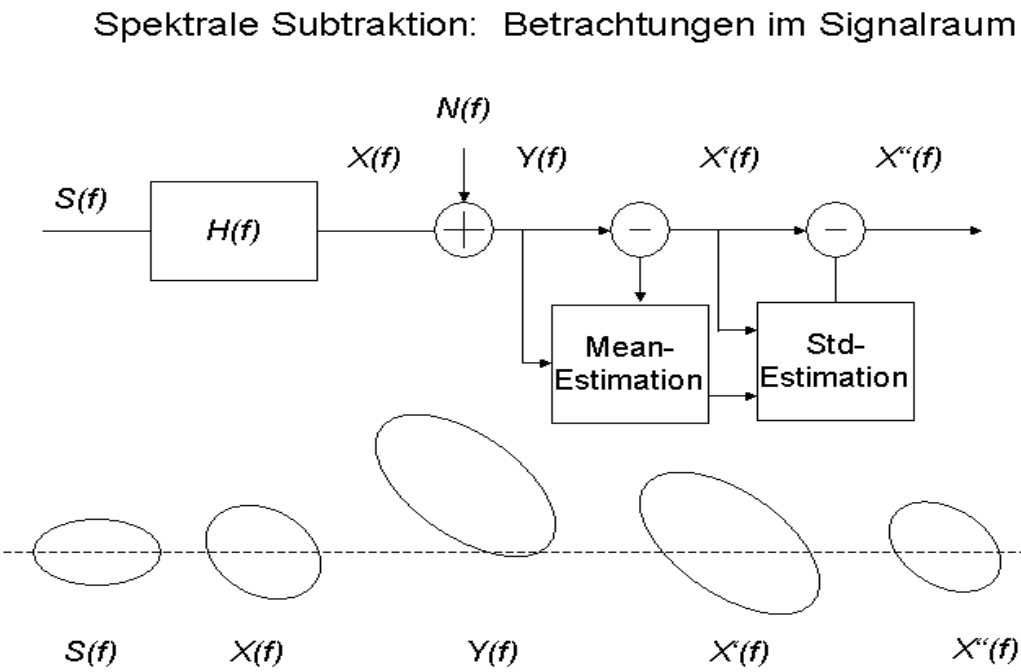

Abbildung 6.1: ***Die Signalverarbeitungskette kann durch eine affine Transformation $\boldsymbol{y = Ax + b}$ beschrieben werden. Mit der Matrix $\boldsymbol{A}$ werden Drehungen, Skalierungen, Scherungen usw. im Signalraum berücksichtigt (Kanaleinflüsse). Dagegen sorgt der Vektor $\boldsymbol{b}$ für Verschiebungen und entsprechende Varianzänderungen im Signalraum. Dabei ist zu beachten, dass sowohl der Kanal als auch das Störsignal als zeitvariante oder auch zeitinvariante Einflüsse vertreten sein können.***

6.2 Verfahren zur Schätzung der Rauschleistungsdichte

Zur Berechnung des störbefreiten Sprachsignals wird eine Schätzung der Rauschleistungsdichte benötigt. Je präziser diese Schätzung ist, desto besser können die Verfahren zur Geräuschreduktion die Geräuschkomponenten ohne störende Artefakte reduzieren. Häufig wurden diese Schätzungen bei Verfahren zur Verbesserung der Sprachqualität verwendet. Unterschätzungen führen hier meist zu unnatürlich klingenden Reststörungen, Überschätzungen dagegen bewirken unnötige Dämpfungen und führen somit zu Verzerrungen des Sprachsignals. Diese Artefakte führen nicht nur zu Qualitätseinbußen bei der Signalverbesserung, sie schränken auch die Leistungsfähigkeit von Spracherkennsystemen ein. Ziel dieses Abschnitts ist es nun Möglichkeiten zur Schätzung der Rauschleistungsdichte darzustellen. Zunächst werden einige bekannte Verfahren vorgestellt, in Abschnitt 6.4 wird dann ein Verfahren vorgestellt, welches die Vorteile der genannten Verfahren miteinander kombiniert.

Bei den hier verwendeten einkanaligen Geräuschreduktionsverfahren besteht die Hauptschwierigkeit darin, eine Schätzung des Leistungsdichtespektrums allein aus dem gemessenen Signal zu ermitteln. In der Literatur findet man im Wesentlichen zwei Klassen von Verfahren, mit denen sich das Problem zufriedenstellend lösen lässt.

Zur ersten Klasse gehören diejenigen Verfahren, welche über einen so genannten ***V**oice-**A**ctivity- **D**etector* zwischen Pause- und Sprachsegmenten unterscheiden. Das Spektrum der Rauschstörung wird hier ausschließlich in den Pauseabschnitten geschätzt und anschließend zur Geräuschunterdrückung verwendet. Bei schlechten SNR müssen hohe Anforderungen an die Leistungsfähigkeit des *VAD* gestellt werden. Darüber hinaus ist dieser Ansatz nur für quasistationäre Störgeräusche geeignet.

Die zur zweiten Klasse gehörenden Verfahren benötigen keinen *VAD*. Diese Verfahren beruhen auf der Beobachtung, dass während einer Sprachaktivität nicht alle Frequenzbereiche Sprachanteile aufweisen müssen. Das als Minimum-Statistik bekannte Verfahren wurde von *[Martin-94]* vorgestellt und wurde seither hinsichtlich Recheneffizienz und Leistungsfähigkeit zum Teil vom Autor selbst aber auch von anderen Autoren *[Doblinger-95]* immer wieder verbessert. Dieses Verfahren verfolgt für jede einzelne Spektralkomponente die Minima des Spektrums des gestörten Signals unabhängig von Sprachaktivität. Es hat jedoch den Nachteil, Schätzwerte mit Bias zu liefern, für den ein Ausgleich notwendig ist. Zur Reduktion der Varianz der Schätzung wurde in *[Martin-01]* eine adaptive Glättung des gemessenen Signals vorgeschlagen, so dass das Schätzergebnis weniger um die mittlere Rauschleistungsdichte schwankt.

Das von *[Hirsch-95]* vorgeschlagene Verfahren gehört ebenfalls zur zweiten Klasse und basiert auf der Verwendung von Kurzzeit-Histogrammen, die für jede einzelne Spektralkomponente erstellt werden. Da bei Sprachaktivität in der Regel höhere Spektralwerte erreicht werden als dies bei sprachfreien Segmenten der Fall ist, kann je ein optimaler Schwellwert bestimmt werden, mit dem zwischen diesen Signalabschnitten unterschieden werden. Das Kurzzeit-Histogramm einer Spektralkomponente kann auch als Überlagerung von zwei einzelnen Histogrammen (Sprache, Rauschen) verstanden werden. Daher führen die Erwartungswerte der einzelnen Kurzzeit-Histogramme unterhalb der Schwellwerte zu einer Schätzung des Rauschspektrums.

Das Verfahren der Quantilenschätzung *[Stahl-00]* ist dem Verfahren nach *[Hirsch-95]* sehr ähnlich. Hier wird für jede Spektralkomponente eine Anzahl von ungeglätteten Werten gespeichert und der Größe nach sortiert. Entnimmt man dieser Liste statt des Minimums einen Wert an einer bestimmten Stelle der Liste, so werden je nach Position z.B. 25% oder 50% der Werte kleiner als der entnommene Wert sein. Diese Elemente nennt man dann Quantil 25 bzw. Qantil 50. Da die großen Spektralelemente von Sprache meist im oberen Bereich der sortierten Liste zu finden sind, ist ein Wert im unteren Bereich $x < 50\%$ mit großer Wahrscheinlichkeit für das Rauschen repräsentativ.

Analysiert man diese Ansätze, so kommt man zu dem Schluss, dass bei dem einen Verfahren in Abhängigkeit vom gemessenen Spektrum Minima geschätzt werden, beim anderen Verfahren dagegen eine obere Grenze für das Rauschsignal festgelegt wird. Eine Kombination dieser Verfahren bestimmt für alle Spektralkomponenten sowohl die untere als auch die obere Grenze der Rauschkomponente signalabhängig. Werte die innerhalb dieses Rauschbandes fallen, können z.B. für die Schätzung des Erwartungswertes herangezogen werden. Darüber hinaus besteht die Hoffnung, mit dem Rauschband auch eine Schätzung der Varianz des Restrauschens zu erhalten.

6.2.1 Das Verfahren der Minimum-Statistik

Das Verfahren der Minimum Statistik *[Martin-94]* beruht auf der Eigenschaft, dass in den einzelnen Spektralkomponenten auch während einer Sprachaktivität nur kurzzeitig Sprachanteile auftreten. Die Grundidee der Minimum Statistik besteht darin, die Minima dieser Spektralwerte innerhalb von zeitlich begrenzten Signalausschnitten zu bestimmen und in einem FIFO-Speicher zu halten. D.h. es liegt zunächst eine Menge von Minima aus aufeinander folgenden Signalabschnitten vor. Wenn die Länge des FIFO-Speichers nun so gewählt wird, dass damit Sprach- und Pausensegmente erfasst werden können, dann können durch eine wiederholte Minimum-Bestimmung die dem Sprachabschnitt zugeordneten Minima ausgeblendet werden.

Um repräsentative Werte für die Minima zu erhalten, ist jedoch eine Glättung der eingehenden Spektralwerte notwendig. Prinzipiell ist das Verfahren der Minimum Statistik daher in der Lage, Änderungen des Rauchspektrums auch während der Sprachaktivitäten zu folgen, d.h. dieses Verfahren eignet sich zur Schätzung von nichtstationärem Rauschen.

Spektrale Analyse:

Zum Zwecke der Modellierung und Analyse wird das bandbegrenzte und mittelwertfreie Sprachsignal *s(n)* einem Rauschsignal mit gaußscher Charakteristik $N(0,\sigma)$ überlagert. Das nunmehr gestörte Signal *y(n)* wird einer diskreten Kurzzeit-Fouriertransformation unterzogen. Die Fensterlänge beträgt *L*-Abtastwerte, die Verschiebung des Analysefensters erfolgt um *R*- Abtastwerte.

$$Y(m,k) = \sum_{l}^{L-1} y(mR+l)\, h(l)\, e^{-2\pi kl/L} \qquad m \in Z, k \in \{0,1,...,L-1\} \qquad (6.2.1)$$

Die spektralen Abtastwerte sind gegenüber den Abtastwerten im Zeitbereich um den Faktor R dezimiert, *m* entspricht daher bezüglich der Zeitachse einem dezimierten Index. Die Wahrscheinlichkeitsdichtefunktion der Betragsquadrate $|Y(m,k)|^2$ ist durch folgenden Ausdruck gegeben *[Martin-94].*

$$f_{|Y(m,k)|^2}(x) = \frac{u(x)}{\sigma_n^2(\lambda,k)+\sigma_s^2(\lambda,k)} e^{-x/\left(\sigma_n^2(\lambda,k)+\sigma_s^2(\lambda,k)\right)} \qquad (6.2.2)$$

Folgt man der minimalen Leistung des gestörten Signals innerhalb eines endlichen Fensters der Länge D, welches groß genug ist um Signalabschnitte größerer Leistung zu überbrücken, kann man eine Schätzung der Rauschleistung erhalten. Um jedoch repräsentative Werte für diese Minima zu erhalten, ist eine Glättung des gestörten Spektrums notwendig:

$$P(\lambda,k) = \alpha P(\lambda-1,k) + (1-\alpha)\,|Y(\lambda,k)|^2 \qquad (6.2.3)$$

In der Praxis verwendet man bei einer Abtastfrequenz von f_s = 8 kHz Fensterlängen, die etwa 100 spektrale Abtastwerte erfassen, dies entspricht einer Dauer von 12.5 ms. Um den Berechnungsaufwand und den Speicherbedarf zu begrenzen, werden die Spektralkomponenten, über welche die Minimum-Suche ausgeführt wird, in U zeitliche Blöcke eingeteilt. Dabei besteht jeder Block aus V-Spektralwerten. Der Ausgang der Minimum-Statistik für einen Kanal bestimmt sich dann als Minimum der Minima aller vollständigen U-Blöcke mit jeweils V-Spektralwerten (6.2.4). Realisiert man dieses Konstrukt als Ringspeicher, so wird für alle V-Spektralwerte der älteste Block durch einen neuen Block ersetzt. D.h. im Takt von jeweils V-Spektralwerten kann aus allen U-Minima das aktuelle Minimum für das Gesamtfenster der Länge $D = U \cdot V$ berechnet werden. Mit diesem adaptiven Verfahren lassen sich daher auch nichtstationäre Störungen verfolgen.

$$P_{\min}^{D} = \min\{Min_1^V, Min_2^V, ..., Min_U^V\} \qquad (6.2.4)$$

In *[Martin-01]* wird ein verbessertes Verfahren vorgeschlagen, mit dem man die Nachteile welche aus einem Glättungsverfahren mit festem Parameter α resultieren, überwinden kann: In sprachfreien Segmenten ist eher eine starke Glättung wünschenswert, in Segmenten in denen Sprache einsetzt, sollte die Glättung nur schwach ausgeprägt sein. Zu starke Glättungen können im zuletzt genannten Fall zu einer Verbreiterung der spektralen Spitzen führen. Da der Ringspeicher der Minima dann vollständig mit Sprachsignalanteilen überdeckt wird, würde es nunmehr keine repräsentativen Werte für das Rauschen enthalten.

Aus diesem Grund wird daher ein zeit- und frequenzabhängiger Glättungsfaktor eingeführt.

$$\alpha_{opt}(\lambda,k)=\frac{1}{1+\left(\bar{\gamma}(\lambda,k)-1\right)^2} \tag{6.2.5}$$

Der Term $P(\lambda-1,k)/\sigma_n^2(\lambda,k)=\bar{\gamma}(\lambda,k)$ kann als ein geglättetes *a-posteriori* SNR betrachtet werden.

$$\gamma(\lambda,k)=\frac{|Y(\lambda-1,k|^2}{\sigma_n^2(\lambda,k)} \tag{6.2.6}$$

Die Minimumschätzung liefert eine Unterschätzung der wahren Rauschleistung, da der Mittelwert $E\{P_{\min}(\lambda,k)\}$ aber proportional zur wahren Rauschleistung $\sigma_n^2(\lambda,k)$ ist, erhält man mit dem folgenden Ausdruck eine biasfreie Schätzung:

$$\hat{\sigma}_n^2(\lambda,k)=\frac{P_{\min}(\lambda,k)}{E\{P_{\min}(\lambda,k)\}_{|\,\sigma_n^2(\lambda,k)=1}} \tag{6.2.7}$$

Mit dieser biasfreien Schätzung kann nun eine spektrale Subtraktion im Frequenzbereich erfolgen.

6.3 Verfahren zur Geräuschunterdrückung im Frequenzbereich

6.3.1 Lineare Spektrale Subtraktion

Die Auswirkungen additiver Störungen auf das Signalspektrum sind bestimmt durch eine Erhöhung von Mittelwert und Varianz. Dem Ansteigen des Mittelwertes kann man durch Subtraktion einer Schätzung des mittleren Rauschspektrums vom gestörten Signalspektrum entgegenwirken. Die Subtraktion kann sowohl auf dem Betragsspektrum als auch auf dem Leistungsspektrum ausgeführt werden.

$$|\hat{X}(f)|^b=|Y(f)|^b-\alpha\,|\overline{N}(f)|^b \tag{6.3.1}$$

Der Proportionalitätsfaktor α steuert den Wert des zu subtrahierenden mittleren Rauschspektrums. Üblicherweise werden Werte α ≥ 1 gewählt. Im Falle α > 1 spricht man von Übersubtraktion. Insbesondere bei schlechtem lokalen SNR besteht die Gefahr, dass negative Schätzungen des rauschgeminderten Spektrums berechnet werden. Um negative Schätzungen zu vermeiden, werden diese mittels Flooring auf positive Werte abgebildet. Darüber hinaus wird in der Literatur über weitere Methoden des Postprocessing berichtet, welche sich mit der Reduktion der Verarbeitungseffekte (Musical Tones usw.) beschäftigen. Dies lässt eine Interpretation als Wiener Filter zu, d.h. die Filterwirkung kann durch eine vom SNR abhängige Dämpfung erklärt werden.

6.3.2 Nichtlineare Spektrale Subtraktion

Nichtlineare Verfahren basieren auf der Verwendung von lokalen Schätzungen des SNR und der Beobachtung, dass insbesondere bei lokal geringem SNR eine Übersubtraktion verbesserte Signalschätzungen hervorbringt.

$$|\hat{X}(f)|^b = |Y(f)|^b - \alpha(SNR(f))\,|\overline{N}(f)|^b \tag{6.3.2}$$

Mit Einführung eines SNR abhängigen Subtraktionsfaktors kann nun lokal eine Übersubtraktion auf diejenigen Frequenzkomponenten angewendet werden, welche nur ein geringes SNR aufweisen. Eine häufig verwendete Form des SNR- abhängigen Subtraktionsfaktors ist durch die folgende Formel gegeben:

$$\alpha(SNR(f)) = \left(1 + \frac{STD(|N(f)|)}{|\overline{N}(f)|}\right) \tag{6.3.3}$$

Der Faktor bewegt sich innerhalb der 2 Extremfälle, determiniertes Rauschen mit Null-Varianz und weißem Rauschen. In der Spracherkennung wird dieser Faktor daher gewöhnlich auf Werte zwischen 1 und 2 festgelegt.

6.4 Verfahren zur Geräuschreduktion im Zeitbereich

Bei den Verfahren zur Geräuschreduktion von Auditiven Modellen im Zeitbereich muss ebenfalls eine Schätzung der Rauschleistung erfolgen. Zunächst erfolgt eine Analyse der Wahrscheinlichkeitsverteilungen für die einzelnen Signalverarbeitungsblöcke nach Abbildung 6.4.1. Anschließend werden unter Verwendung der in diesem Abschnitt gefundenen Beziehungen die Lineare Subtraktion in Abschnitt 6.4.1 und die Nichtlineare Subtraktion in Abschnitt 6.4.2 behandelt.

Den Ausgangspunkt für beide Verfahren stellt die Überlegung dar, dass Rauschen nicht nur das auditive Spektrum beeinflusst sondern auch die phonetische Kontrastverstärkung der Adaptionsstufe reduziert und den dynamischen Bereich der Haarzelle vermindert. Daher wird von den Autoren in *[Veerecken-95]* der Einsatz eines Algorithmus vorgeschlagen, welcher das Rauschen bereits vor dem Eintritt in die Adaptionsstufe unterdrückt.

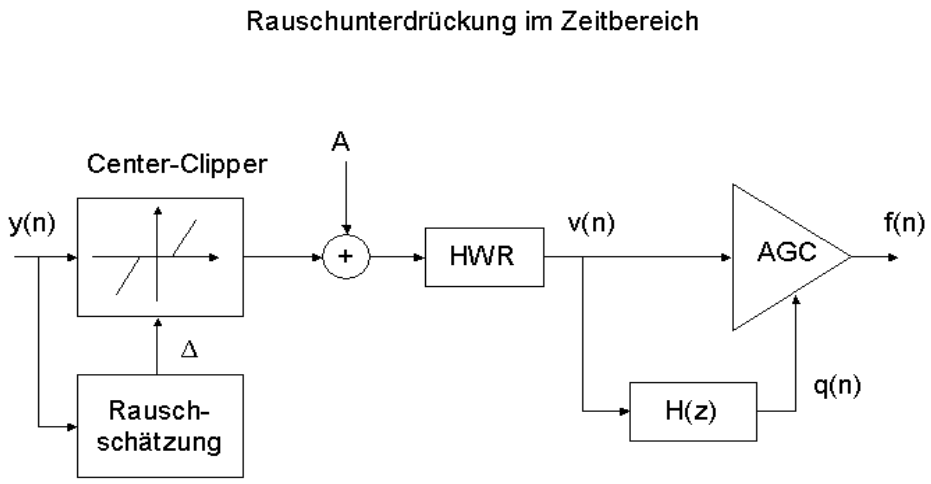

Abbildung 6.4.1 ***Rauschunterdrückung im Zeitbereich: Bei Verwendung des Martens-Immerseel Haarzellen-modells arbeiten alle Module mit der vollen Abtastrate, das gilt auch für die hier nicht dargestellte Filterbank.***

Das Verfahren nach *[Veerecken-95]* arbeitet auf allen Filterausgängen mit der vollen Abtastrate. Bevor die Teilbandsignale dem jeweiligen Haarzellen Modell zugeführt werden, durchlaufen die Teilbandsignale eine Rauschunterdrückung (realisiert durch den CC-Block).

$$y(n) = \begin{cases} x(n) - \Delta(n) & x(n) > \Delta(n) \\ 0 & \text{wenn} \quad |x(n)| \leq \Delta(n) \\ x(n) + \Delta(n) & x(n) < \Delta(n) \end{cases} \tag{6.4.1}$$

Die Idee bei diesem Verfahren ist es, die Standardabweichung σ des Teilbandrauschens zu schätzen und $\Delta(n)$ auf 2σ zu setzen und damit etwa 95 % des Rauschanteils zu unterdrücken.

Später wurde von den Autoren ein nichtlineares Verfahren vorgestellt, welches $\Delta(n)$ adaptiv aus einer Kennlinie schätzt. Zur Bestimmung des optimalen Wertes $\Delta(n)$ wird das nachfolgende Modell zur Verteilung der Störgröße angenommen.

$$N(0,\sigma) = f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}} \tag{6.4.2}$$

Die Verteilung der Signalwerte für einen Kanal nach der Gammatone-Filterbank und dem CC-Block berechnet sich wie folgt *[Vereecken-95]*:

$$f_Y(y) = F_X(|x(n)| \leq \Delta)\delta(y) + f_X(y+\Delta)u(y) + f_X(y-\Delta)u(-y) \tag{6.4.3}$$

Dabei wird die Notation $\delta(.)$ für die Dirac-Funktion und die Notation $u(.)$ für die Einheitssprung-Funktion verwendet. Die Verteilung setzt sich aus drei Anteilen zusammen (Anhang A):

- der kummulativen Wahrscheinlichkeitsverteilung für Signalwerte, die betragsmäßig kleiner als die Schwelle $\Delta(n)$ sind (diskreter Anteil)
- der Wahrscheinlichkeitsverteilung für Signalwerte, die oberhalb der positiven Schwelle $\Delta(n)$ liegen (erster kontinuierlicher Anteil)
- der Wahrscheinlichkeitsverteilung für Signalwerte, die unterhalb der positiven Schwelle $\Delta(n)$ liegen (zweiter kontinuierlicher Anteil)

D.h. die Verteilung setzt sich aus zwei kontinuierlichen und einem diskreten Anteil zusammen (siehe Anhang A). Die Wahrscheinlichkeit, dass der Betrag eines Teilbandsignals kleiner als Δ ist, stellt das Gewicht für den diskreten Anteil dar und ist durch (6.4.4) gegeben:

$$F_X(|x(n)| \leq \Delta) = erf\left(\frac{\Delta}{\sigma\sqrt{2}}\right) \tag{6.4.4}$$

Die Definition von $erf(z)$ ist durch (6.4.5) gegeben (siehe auch Anhang A):

$$erf(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2} dt \tag{6.4.5}$$

Für die Dichteverteilung der Zufallsgröße *V* (nach der Halbwellen-Gleichrichtung) erhält man den folgenden Ausdruck:

$$f_V(v) = f_Y(v-A)u(v) + F_Y(y(n) \leq -A)\delta(v) \tag{6.4.6}$$

Bildet man den Erwartungswert, liefern nur diejenigen Werte der Verteilung $f_Y(v-A)$ einen Beitrag, die auf der positiven *v*-Achse liegen.

$$E\{V\} = \int_0^\infty v f_Y(v-A)dv$$

$$= \frac{A-\Delta}{2} + \frac{A+\Delta}{2}\Phi(\gamma) = E\{q(n)\} \tag{6.4.7}$$

Dabei ist zu beachten, dass eine Schätzung von *E{q(n)}* am Tiefpassfilter abgegriffen werden kann. Für die Gleichung (6.4.7) gelten folgende Beziehungen *[Vereecken-95]*:

$$\gamma = \frac{A+\Delta}{\sigma\sqrt{2}} \qquad \text{und} \qquad \Phi(\gamma) = \frac{1}{\sqrt{\pi}}\frac{e^{-\gamma^2}}{\gamma} + erf(\gamma) \tag{6.4.8}$$

6.4.1 Lineare Subtraktion

Das lineare Verfahren wird in *[Veerecken-95]* beschrieben und ist unter dem Namen ***L**inear **N**oise **M**agnitude **S**ubtraction* bekannt. Zur Steuerung der Verstärkung (AGC) des Haarzellen Modells wird das Ausgangsignal *q(n)* des Tiefpassfilters verwendet. Das Minimum dieses Signals ($q_{\min}(n)$) dient dabei als Schätzwert für den Erwartungswert von *V*.

$$\gamma(n) = \Phi^{-1}(\phi) \qquad \text{mit} \quad \phi = \frac{2q_{\min}(n) - A + \Delta_{old}}{A + \Delta_{old}} \tag{6.4.9}$$

Die Standardabweichung σ zum Zeitpunkt *n* ergibt sich mit (6.4.8) und (6.4.9) zu:

$$\sigma(n) = \frac{A+\Delta}{\gamma(n)\sqrt{2}} \tag{6.4.10}$$

Mit Schätzung der Standardabweichung erhält man für den Schwellwert $\Delta(n) = 2\sigma(n)$, mit diesem Wert wird der Center Clipper gesteuert.

6.4.2 Nichtlineare Subtraktion

Neben dem linearen Verfahren wurde von den Autoren auch ein nichtlineares Verfahren entwickelt entwickelt und in *[Vereecken-96]* vorgestellt, ***N**onlinear **N**oise **M**agnitude **S**ubtraction*. Die Motivation nichtlineare Verfahren zur Rauschunterdrückung zu verwenden, liegt darin begründet, dass der Einsatz nichtlinearer Verfahren insbesondere bei schlechtem SNR bezüglich der Erkennrate den linearen Verfahren überlegen sind. Die zugrunde liegende Idee ist es, bei guten SNR weniger Rauschen zu subtrahieren und bei schlechtem SNR dagegen die Subtraktion zu verstärken. Zur Berechnung der charakteristischen Kennlinie wird der folgende zweistufige Ansatz verfolgt:

$$q_{\min} = \frac{A-\Delta}{2} + \frac{A+\Delta}{2}\Phi_1(\gamma) = A + \frac{\sigma}{\sqrt{2}}\Phi_2(\gamma) \tag{6.4.11}$$

Dabei gelten nun folgende Beziehungen:

$$\Phi_1(\gamma) = \Phi(\gamma) \qquad \text{und} \qquad \Phi_2(\gamma) = \gamma[\Phi_1(\gamma) - 1] \tag{6.4.12}$$

Entsprechend dem linearen Verfahren gilt unverändert:

$$\gamma(n) = \Phi_1^{-1}(\phi_1) \quad \text{mit} \quad \phi_1 = \frac{2q_{\min} - A + \Delta}{A + \Delta_1}$$

$$\text{und} \qquad \sigma(n) = \frac{A + \Delta_1}{\gamma_1(n)\sqrt{2}} \tag{6.4.13}$$

Nachdem die Standardabweichung σ vorliegt wird $\Delta(n)$ so gewählt, dass der Erwartungswert von *V* bzw. $q_{\min}(n)$ für sprachfreie Segmente konstant bleibt.

$$q_{\min} = A + \frac{\sigma}{\sqrt{2}}\Phi_2(\gamma) = A + \varepsilon A \tag{6.4.14}$$

Mit (6.4.14) erhält man

$$\Phi_2(\gamma_2) = \frac{\varepsilon A\sqrt{2}}{\sigma} \quad \text{bzw.} \quad \gamma_2 = \Phi_2^{-1}(\phi_2)\,, \tag{6.4.15}$$

so dass Δ_2 nach (6.4.15) und (6.4.10) berechnet werden kann:

$$\Delta_2(n) = \sigma(n)\gamma_2(n)\sqrt{2} - A \tag{6.4.16}$$

bzw.

$$\Delta_2(n) = \max[\sigma(n)\gamma_2(n)\sqrt{2} - A, 0] \tag{6.4.17}$$

Im Gegensatz zum linearen Verfahren wird die Standardabweichung nun nicht mehr mit dem konstanten Faktor 2 sondern mit dem Clipping Faktor $\gamma_2(n)\sqrt{2}$ multipliziert. Trägt man den Clipping Faktor gegen die Standardabweichung auf, erhält man eine nichtlineare Kennlinie, welche die oben beschriebene Forderung bezüglich verschiedener SNR einhält.

6.5 Geräuschunterdrückung für Auditive Modelle im *RMS*-Bereich

Das hier beschriebene Verfahren berücksichtigt nun die Implementierung des Haarzellenmodells im Frequenz- bzw. im *RMS*-Bereich. Um Haarzellenmodelle wirksam in die Signalverarbeitungskette integrieren zu können, ist es notwendig den dynamischen Bereich der Haarzelle an den gegebenen Rauschpegel anzupassen. Aufgrund der Überlagerung von Rausch- und Sprachsignal führen große Verstärkungen des Rauschsignals zu einem Sättigungsverhalten des Gesamtsignals. Dies hat zur Folge, dass die Informationsübertragung durch die Haarzelle stark beeinträchtigt wird. In *[Perdigao-99]* wird daher ein adaptives Verfahren vorgeschlagen, welches den Rauschanteil des Signal V_i in Abhängigkeit von der gemessenen Rauschleistung absenkt und an die Maskierungsschwelle *A* klemmt, d.h. der gesamte dynamische Bereich von ~ 65 dB steht der Übertragung des Sprachsignals zur Verfügung. Mit anderen Worten: das Ziel des Verfahrens besteht darin, in sprachfreien Segmenten den Erwartungswert des Signals *V* für alle Teilbänder *i* konstant zu halten.

Das folgende Verfahren basiert auf dem bereits im vorangegangenen Abschnitt beschriebenen nichtlinearen Verfahren im Zeitbereich. Dabei wurde für jedes Filterband die Standardabweichung σ geschätzt und aus der für dieses Band charakteristischen Kennlinie adaptiv der Wert für $\Delta(n)$ bestimmt. Für die Verteilung der Rauschamplituden wurde $N(0,\sigma)$ angenommen.

Im *RMS*-Bereich verliert diese Annahme über die Verteilung ihre Gültigkeit, daher werden zunächst die Verteilungen der Zufallsgrößen Y_i gemessen. Als Eingangsgröße diente mittelwertfreies weißes Rauschen mit einer Standardabweichung von σ = 100. Wie Abbildung 6.5.1 zeigt, können die gemessenen Verteilungen der einzelnen Kanäle durch die Gammaverteilung angenähert werden.

$$f_X(x) = \frac{c^{b+1}}{\Gamma(b+1)} x^b e^{-cx} u(x) \tag{6.5.1}$$

Die Gammaverteilung kann durch die Parameter *b* und *c* beschrieben werden. Für den Mittelwert und die Varianz gelten folgende Zusammenhänge:

$$\mu_x = \frac{b+1}{c}, \quad \sigma^2 = \frac{b+1}{c^2} = \frac{\mu_x}{c} \tag{6.5.2}$$

Da im *RMS*-Bereich ausschließlich positive Werte auftreten, kann die Maskierung durch den folgenden Ausdruck berücksichtigt werden.

$$V_i[m] = \max(Y_i[m] - \Delta, 0) + A \tag{6.5.3}$$

In *[Perdigao-99]* konnte durch Experimente gezeigt werden, dass für jeden Kanal *i* ein Parameter b_i geschätzt werden kann, welcher unabhängig von der Varianz σ^2 des Rauschsignals ist. Der Parameter c_i ist nach (6.5.2) umgekehrt proportional zu $E\{Y_i\}$ und demnach auch abhängig von der Standardabweichung σ.

Mit der funktionalen Beschreibung der Verteilung der Zufallsgrößen Y_i gelingt es nun den Erwartungswert $E\{V_i\}$ analytisch zu erfassen. Zunächst ergibt sich für die Verteilungsfunktion von V_i:

$$f_{V_i}(V_i) = F_{Y_i}(y_i \le \Delta)\delta(V_i - A) + f_{V_i}(V_i - (A - \Delta))u(V_i - A) \tag{6.5.4}$$

mit den Parametern aus (6.5.2) kann der Erwartungswert von V_i wie folgt berechnet werden:

$$E\{V_i\} = A - \Delta(1 - G(c_i\Delta)) + \frac{b_i + 1}{c_i}(1 - G(c_i\Delta, b_i + 1)) \tag{6.5.5}$$

In (6.5.5) wird die unvollständige Gammafunktion *G(x,b)* verwendet, diese ist folgt definiert:

$$G(x,b) = \frac{1}{\Gamma(b+1)} \int_0^x t^b e^{-t} dt \tag{6.5.6}$$

bzw.

$$G(x,b+1) = G(x,b) - \frac{x^{b+1} e^{-x}}{\Gamma(b+2)} \tag{6.5.7}$$

Mit den Parametergleichungen aus (6.5.2) und den beiden obigen Gleichungen erhält (6.5.5) die folgende Form:

$$E\{V_i\} = A + E\{Y_i\}\left((1 - \Delta / E\{Y_i\})(1 - G(c_i\Delta, b_i) + \frac{(c_i\Delta)^{b_i+1} e^{-c_i\Delta}}{\Gamma(b_i + 2)} \right) \tag{6.5.8}$$

An dieser Stelle liegt eine Beschreibung des Erwartungswertes von *V* im Teilband *i* vor. Das Ziel des Verfahrens besteht darin, den 2. Term in (6.5.8) konstant zu halten. Um das weitere Vorgehen transparent zu halten, wird dieser Term einer genaueren Betrachtung unterworfen:

Der Erwartungswert $E\{Y_i\}$ des Teilbands *i* ist proportional zur Standardabweichung σ des Rausch-signals. Der Parameter c_i ist sowohl vom Kanal als auch von der Standard-abweichung σ des Rauschsignals abhängig. Der Parameter b_i ist nicht von σ abhängig, liegt aber für jeden Kanal in unterschiedlichen Ausprägungen vor.

D.h. als nächstes kann vor dem Hintergrund der Proportionalitätsbeziehungen der Parameter c_i eliminiert werden. In diesem Sinne wird zunächst der Ausdruck $\Delta / E\{Y_i\}$ ersetzt:

$$\frac{\Delta}{E\{Y_i\}} = \frac{\Delta}{\sigma}\frac{\sigma}{E\{Y_i\}} = r\frac{\sigma}{E\{Y_i\}}. \tag{6.5.9}$$

Mit der Substitution $r = \frac{\Delta}{\sigma}$ gilt für $c_i\Delta$ mit (6.5.2):

$$c_i\Delta = (b_i + 1) r \frac{\sigma}{E\{Y_i\}} \tag{6.5.10}$$

Aufgrund der Proportionalität kann die Gleichung (6.5.8) verkürzt in folgender Form dargestellt werden

$$E\{V_i\} = A + \sigma\Phi_i(r) \tag{6.5.11}$$

Mit den Gleichungen (6.5.11), (6.5.9) und (6.5.8) kann nun die Funktion $\Phi_i(r)$ nach Gleichung (6.5.12) berechnet werden:

$$\Phi_i(r) = \frac{E\{Y_i\}}{\sigma}\left((1 - r\frac{\sigma}{E\{Y_i\}})(1 - G((b_i+1)r\frac{\sigma}{E\{Y_i\}}, b_i)) + \frac{\left((b_i+1)r\frac{\sigma}{E\{Y_i\}}\right)^{b_i+1} e^{-\left((b_i+1)r\frac{\sigma}{E\{Y_i\}}\right)}}{\Gamma(b_i+2)} \right)$$

Diese Funktion hängt nicht von c_i ab und liefert an der Stelle $q=0$ die gesuchte Proportionalitätskonstante:

$$\Phi_i(0) = \frac{E\{Y_i\}}{\sigma} \tag{6.5.13}$$

Nun kann für $r = (\Delta / \sigma)$ ein sinnvolles Intervall vorgegeben und die gefundene Konstante $\Phi_i(0)$ in (6.5.12) eingesetzt werden. Um $E\{V_i\}$ in sprachfreien Segmenten konstant zu halten, setzt man den 2. Term in (6.5.12) einfach auf Eins:

$$\frac{1}{\sigma} = \Phi_i\left(\frac{\Delta}{\sigma}\right) \tag{6.5.14}$$

Mit (6.5.12) und (6.5.13) erhält man dann für jedes Teilband das Verhältnis $\Delta/E\{Y_i\}$, als Funktion von $E\{Y_i\}$:

$$\frac{\Delta}{E\{Y_i\}} = \frac{1}{\Phi_i(0)} \Phi i^{-1}\left(\frac{\Phi_i(0)}{E\{Y_i\}}\right) \tag{6.5.15}$$

Die Multiplikation einer Schätzung von $E\{Y_i\}$ mit diesem Verhältnis liefert dann für die Subtraktion in (6.5.3) den gesuchten Wert Δ. Die Schätzwerte $E\{Y_i\}$ können entweder in Pausenabschnitten oder basierend auf dem Verfahren der Minimum- Statistik gewonnen werden. Die Wertepaare $\Delta/E\{Y_i\}$, $E\{Y_i\}$ werden in einer Tabelle abgelegt, so dass auf Δ effizient zugegriffen werden kann.

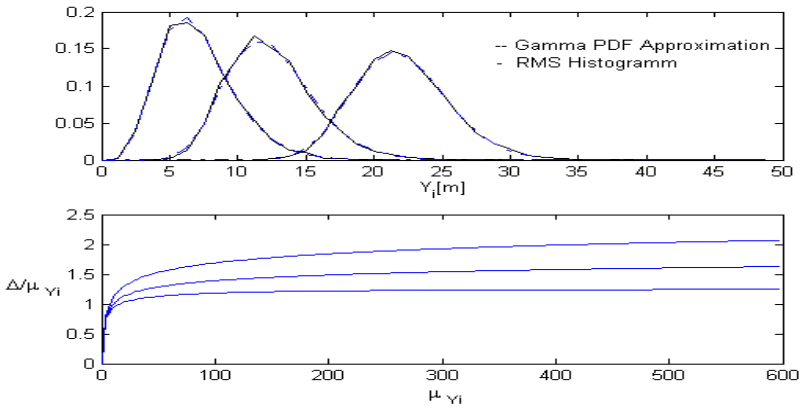

Abbildung 6.5.1: ***Darstellung des Histogrammverlaufs und Approximation der Verteilungsdichten durch die Gammafunktion für die RMS-Kanäle i=1,17 und 35. Darunter ist der Verlauf der dazugehörigen nichtlinearen Kennlinien zur Bestimmung des Wertes Δ_i gegeben.***

Zusammenfassung zur Berechnung der charakteristischen Kennlinien:

- Berechnung der Histogramme für eine gegebene Standardabweichung σ
- Approximation der Histogramme durch die Gammaverteilung, Schätzung der Parameter c_i und b_i
- Berechnung von $\Phi_i(r)$, Bestimmung der Proportionalitätskonstanten
- Erstellung der Tabelle für die Wertepaare ($\Delta/E\{Y_i\}$, $E\{Y_i\}$)

Dieses Verfahren kann nun für alle Modelle verwendet werden, welche im *RMS*-Bereich angesiedelt sind. Es müssen lediglich alle Schritte, welche in der Zusammenfassung dargelegt worden sind, für die einzelnen *RMS*-Teilbänder der entsprechenden Modelle wiederholt werden.

7 Vereinheitlichtes Auditives Modell mit Geräuschunterdrückung

In den vorangegangenen Abschnitten wurden theoretische auditive Modelle, deren Implementierungen und verschiedene Verfahren zur Rauschunterdrückung vorgestellt. Aus einer abstrakteren Sicht zeigt sich, dass aus dieser Vielfalt von Modellen und Methoden Kernmodule abgeleitet werden können, deren Beziehungen zueinander ein vereinheitlichtes integratives auditives Modell darstellen. Dieses abstraktere Modell berücksichtigt zunächst die folgenden Bestandteile von auditiven Modellen:

- Gehörrichtige Filterbank
- Modulationsaspekte
- Nichtlineare Kompression
- Dekorrelation (*Independent Component Analysis*)

Darüber hinaus enthält das Modell im Kern nur ein einziges Verfahren zur Rauschminderung, dieses kann sowohl im Zeit- als auch im Frequenzbereich verwendet werden.

Mit dieser Abstraktion kann die Komplexität und die Vielfalt der Modelle soweit reduziert werden, dass sich aus ingenieurstechnischer Sicht effiziente Implementierungen finden lassen. D.h. das Modell enthält im Wesentlichen nur noch einige wenige abstrakte Module, die die oben angeführten Bestandteile enthalten, aber diese Bestandteile können in unterschiedlicher Art und Weise implementiert worden sein. Dadurch lassen sich die Eigenschaften des Modells bei unterschiedlicher Realisierung der Module untersuchen. Mit den unterschiedlichen Realisierungen der Module können sowohl klassische *VVM* (cepstrale Darstellung) als auch reine auditive *VVM* untersucht werden. Mit dieser Vorgehensweise wird ein Vergleich der verschiedenen *VVM* bereits auf der Merkmalsebene möglich. Auf dieser Ebene lässt sich beispielsweise die Zuverlässigkeit oder Robustheit der verschiedenen *VVM* beurteilen.

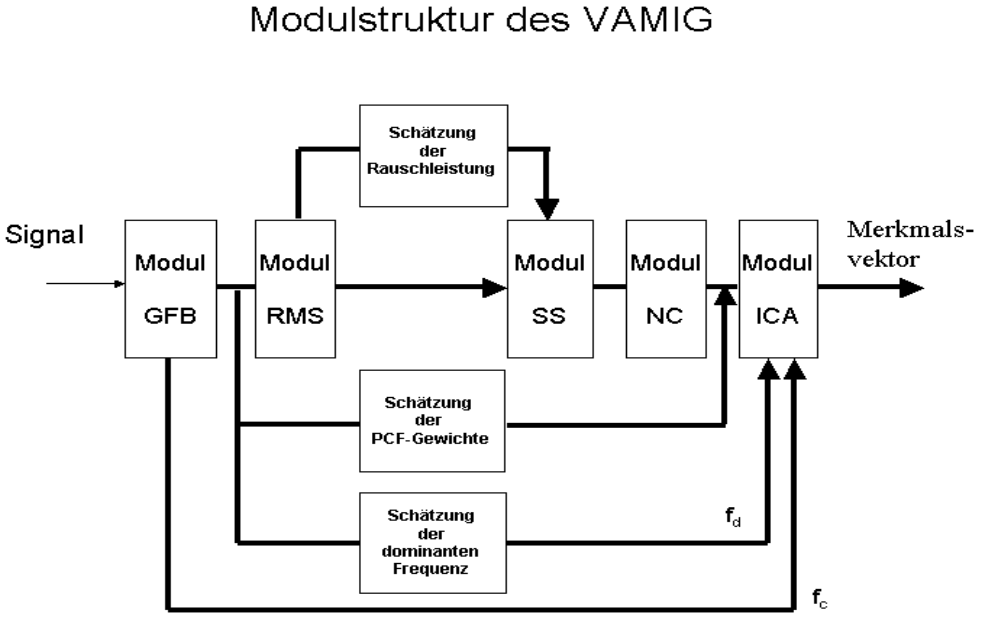

Abbildung 7.1 ***Module der Vorverarbeitungsmethoden in der Signalverarbeitungskette. Die einzelnen Module enthalten jeweils unterschiedliche Realisierungsvarianten bestimmter Verarbeitungselemente. Einige Elemente wie der DF-Estimator kommen nur bedingt zur Anwendung. Der integrative Charakter betrifft die Einbeziehung der Spektralen Subtraktion in das Modell. Die Vereinheitlichung bezieht sich auf den eigentlichen Zweck der einzelnen Module. Die Fensterung kann dem Modul GFB zugeschlagen werden, im Frequenzbereich findet sie vor der FFT statt, im Zeitbereich erst nach der Berechnung der Teilbandsignale.***

7.1 Modul Gehörrichtige Filterbank

Die gehörrichtige Filterbank wird in dieser Arbeit gemäß der Artikulationstheorie durch 3 verschiedene Realisierungen mit jeweils 21 bzw. 39 Teilbändern repräsentiert, alle Realisierungen arbeiten mit einer Abtastfrequenz von 8000 Hz:

- Gammatone- Filterbank (im Frequenzbereich)
- Shamma- Filterbank (im Zeitbereich, asymmetrische Filter)
- Waveletpaket- Zerlegung (im Zeitbereich mit Mehrfachauflösung)

Der Zweck dieses Moduls besteht in der Zerlegung des Signals in ein Multibandsignal. Die Multibandsignale können entweder als Vollbandsignale (Gammatone) oder echte Teilbandsignale (Wavelets, Shamma) auftreten.

7.2 Modul *RMS*-Berechnung

Die Berechnung der *RMS*-Werte kann - wie weiter oben gezeigt - über die Teilbandenergien sowohl im Frequenzbereich als auch im Zeitbereich erfolgen. Darüber hinaus können die *RMS*- Werte auch aus den Modulationsenergien der Teilbänder (*TEO*) im Zeitbereich ermittelt werden. Im Zeitbereich (Shamma, Wavelet) wird die Fensterung nach der Filterbank entsprechend dem Schema in Kapitel 5 vorgenommen. Im Frequenzbereich erfolgt die Fensterung vor der Filterbank. Die Fensterlänge beträgt für alle Varianten 32 ms, die Verschiebung des Fensters beträgt ebenfalls für alle Varianten 10 ms. Mit diesem Modul wird die Erregungsstärke in den einzelnen Teilbändern für jeden einzelnen Frame bestimmt. Im Falle des *TEO* dagegen, wird die Stärke der Modulationsenergien für jedes einzelne Teilband berechnet, dies ist für den Frequenzbereich nicht realisiert und stellt somit die einzige Abweichung im Konzept des *VAMIG* dar. Zur besseren Übersicht werden die Vorschriften zur Berechnung des *RMS* nochmals zusammengefasst:

- Berechnung im Frequenzbereich gemäß (4.2.3):

$$y_i^{RMS} = \sqrt{y_i^P} = \frac{1}{N}\sqrt{\sum_N |X_i(k)|^2} \tag{7.2.1}$$

- Berechnung im Zeitbereich basierend auf der Parsevalschen Beziehung:

$$y_i^{RMS} = \sqrt{y_i^P} = \sqrt{\frac{1}{N_i}\sum_{N_i} x_i^2(n)} \tag{7.2.2}$$

- Berechnung im Zeitbereich basierend auf dem *TEO* gemäß (3.2.10):

$$y_i^{RMS} = \sqrt{\frac{1}{N_i}\sum_{N_i} \Psi[x_i(n)]} \tag{7.2.3}$$

7.3 Modul Spektrale Subtraktion

Für die Minderung der Rauscheinflüsse kommt einerseits das Verfahren der Nichtlinearen Subtraktion nach Perdigao zur Anwendung, andererseits wird die Kombination Lineare Subtraktion des Mittelwertes des geschätzten Rauschsignals und Subtraktion der Standardabeichung des geschätzten Rauschsignals verwendet. Die Schätzung des Rauschsignals basiert auf der Methode der Minimum-Statistik. Dieses Verfahren wird sowohl im Frequenzbereich als auch im Zeitbereich angewendet. Zudem findet es Anwendung bei den *RMS*-Werten welche auf dem *TEO* basieren (Waveletzerlegung).

7.4 Modul Nichtlineare Kompression

In diesem Modul wird die statische Kompression über den Logarithmus der *RMS*-Werte realisiert. Das Haarzellenmodell nach Vereecken bzw. Perdigao entspricht der dynamischen nichtlinearen Kompression. Wie bereits weiter oben beschrieben, kann man sich vom Haarzellenmodell eine erhöhte Robustheit und bessere Segmentierungseigenschaften erwarten. Dieses Modul dient im Wesentlichen der Kompression des Wertebereichs der Erregungsstärken.

7.5 Modul Dekorrelation

In diesem Modul kann einerseits die klassische Variante also eine *DCT* zur Anwendung kommen:

$$ICA_i = \sum_{l=1}^{L} \log(NC(l)) \cos\left[\frac{i(l-0.5\pi)}{L}\right] \qquad (7.5.1)$$

Alternativ steht die Methode *LIN* zur Verfügung. Hier besteht die Aufgabe des Moduls in der kompakten und redundanzfreien Darstellung des Signals. Der entscheidende Unterschied zur *DCT* besteht darin, dass das *LIN* lokal operiert.

Lokale Störungen (entstehen zum Beispiel beim Flooring im Modul Spektrale Subtraktion) verteilt die *DCT* auf den gesamten cepstralen Vektor. Dies führt zu deutlichen Abweichungen gegenüber einem ungestörten cepstralen Vektor. Dies widerspricht der Modellierung entsprechend der Artikulationstheorie, diese fordert eine unabhängige Teilbandverarbeitung.

Auf die Realisierung des *LIN* wird im Folgenden etwas näher eingegangen. Das hier verwendete *LIN* entspricht im Kern dem *LIN,* welches bereits durch das *SBCOR*-Verfahren beschrieben wurde. Der Gewichtsfaktor welches das *LIN* für jeden Kanal liefert, hängt nach (7.5.5) vom Abstand zwischen der dominanten Frequenz und der Mittenfrequenz des jeweiligen Kanals ab. Für die Gammatone-Filterbank im Frequenzbereich kann die dominante Frequenz für jeden Frame durch den Schwerpunkt des jeweiligen Kanals bestimmt werden.

$$f_d^i = \frac{\int_0^{F_s/2} f\,|H_i(f)|^2 |X(f|^2\,df}{\int_0^{F_s/2} |H_i(f)|^2 |X(f|^2\,df} \qquad (7.5.2)$$

Im Zeitbereich (Shamma-Filterbank) führt die Bestimmung der Nulldurchgangsrate auf die dominante Frequenz.

Der Ort des Nulldurchgangs wird durch lineare Interpolation zwischen denjenigen aufeinander folgenden Abtastwerten bestimmt, für die das Produkt der beiden Abtastwerte negativ ist. Bezeichnet man die Zeitpunkte der Nulldurchgänge mit t_k und die Amplitude zwischen zwei Nulldurchgängen mit A_k dann liefert (7.5.3) eine Schätzung der mittleren halben Periodendauer für ein festes Zeitfenster in dem N_Z Nulldurchgänge gemessen wurden:

$$\tilde{T}_H^i = \frac{\sum_{k=1}^{N_z^i-1} |A_k^i|^2 (z_{k+1} - z_k)}{\sum_{k=1}^{N_z^i-1} |A_k^i|^2} \tag{7.5.3}$$

Um eine größere Robustheit gegenüber rauschartigen Störungen zu gewährleisten, wird allerdings ähnlich wie in *[Kim-99]* eine Bewertung der Abstände zwischen den Nulldurchgängen durch die entsprechenden Amplituden vorgenommen. Höhere Amplituden, d.h. energiereiche Komponenten werden in (7.5.3) stärker gewichtet als energiearme Komponenten. Die dominante Frequenz im Kanal *i* erhält man letztlich durch die Berechnung von (7.5.4).

$$f_d^i = \frac{1}{2\tilde{T}_H^i} \tag{7.5.4}$$

Bei den Waveletpackets sind zusätzlich die jeweils unterschiedlichen Abtastfrequenzen in den einzelnen Teilbändern zu beachten. Mit der dominanten Frequenz und der zugehörigen Mittenfrequenz kann nun gemäß (7.5.5) ein Gewichtsfaktor berechnet werden. Diese Formel entspricht der in Abschnitt 3 entwickelten Beziehung zur Gewichtung der Filterkurven des *SBCOR*-Verfahrens.

$$W_i = \frac{(1-\alpha)(\cos 2\pi f_d^i / f_c^i - \alpha)}{1 - 2\alpha \cos 2\pi f_d^i / f_c^i + \alpha^2} \tag{7.5.5}$$

Mit dem Parameter α kann die Frequenzauflösung bzw. die spektrale Schärfe modifiziert werden. Werte von α nahe 1 führen zu einer hohen Frequenzauflösung (breites Fenster im Zeitbereich) α nahe 0 führt dagegen zu einer hohen spektralen Schärfung.

Befindet sich die dominante Frequenz in der Nähe der Mittenfrequenz, so strebt der Gewichtsfaktor seinem Maximalwert entgegen. Umso weiter sich die dominante Frequenz von der Mittenfrequenz entfernt, desto geringer wird der Gewichtsfaktor. Am Ende der Verarbeitungskette wird im Modul *LIN* dieser Faktor mit dem entsprechenden Ausgangswert des NC-Moduls im jeweiligen Kanal multipliziert und bestimmt somit das auditive Spektrum.

$$ICA_i = NC_i W_i \tag{7.5.6}$$

Je nach Auswahl der Realisierungsvarianten können nun z.B. klassische cepstrale Merkmale generiert werden: Gammtone Filterbank + *RMS* + *LOG* + *DCT*, es können aber auch reine auditive Merkmale erzeugt werden, die vollständig der Artikulationstheorie entsprechen: Wavelet- oder Shamma Filterbank + *RMS* + *HC* + *LIN*. Neben diesen beiden Realisierungen besteht noch die Möglichkeit, hybride *VVM* zu untersuchen.

Das *VAMIG* kann auch als ein Modell mit variabler Parametrierung verstanden werden. Die Parametrierung geschieht durch die Auswahl der Realisierungsformen der einzelnen Module. Dementsprechend erzeugt das Modell jeweils qualitativ unterschiedliche Merkmale. Im Hinblick auf einen möglichen Paradigmenwechsel sollen nun diejenigen Merkmale ausgewählt werden, welche über einen weiten Bereich unterschiedlicher Störsituationen und SNR-Variationen möglichst robust erscheinen. Die Methode, mit der robuste Merkmale und damit die dazugehörige Parametrierung des abstrakten Modells ermittelt werden kann, wird im nächsten Kapitel vorgestellt.

8 Untersuchungen zur Robustheit

Der Begriff der Robustheit ist zwar ein weitläufig verwendeter Begriff, gleichzeitig handelt es sich aber auch um einen sehr unscharfen Begriff. In diesem Kapitel werden nun zwei konkrete Definitionen der Robustheit herausgearbeitet, mit denen die quantitative Bewertung der Robustheit einer *VVM* ermöglicht wird. Das Ziel dieser Anstrengung besteht darin, unterschiedliche *VVM* bereits auf der Merkmalsebene miteinander vergleichen zu können.

Die erste Definition lässt sich aus der Informationstheorie theoretisch ableiten und basiert auf dem Begriff der relativen Transinformation. Streng genommen gilt dieses Robustheitsmaß jedoch nur für stationäre weiße Rauschquellen und hat daher vor allem einen theoretischen Wert. Die informationstheoretische Deutung der Robustheit zielt zunächst nur auf eine konkrete Vorstellung des Robustheitsbegriffs ab, für die experimentellen Untersuchungen wird diese Definition nicht berücksichtigt.

Ein zweites Konzept zur Bestimmung der Zuverlässigkeit von Sprachmerkmalen lässt sich aus der Artikulationstheorie von H. Fletcher ableiten. Im Rahmen dieser Theorie entstand ein Artikulationsmodell, mit dem - ausgehend von den SNR in den einzelnen Teilbändern - die so genannte Phonartikulation berechnet werden kann. Die Phonartikulation lässt sich unter Verwendung eines Multibandmodells berechnen. Dabei zeigt die mathematische Struktur des Multibandmodells eine interessante Analogie zur Berechnung der Zuverlässigkeit von Parallelsystemen.

8.1 Informationstheoretische Definition der Robustheit

Intuitiv erscheint ein solches Maß sinnvoll, mit dem ein Abstand von ungestörten Merkmalen zu gestörten Merkmalen angegeben werden kann. Ein solches Maß stellt bspw. die relative Transinformation dar, mit ihr kann gemessen werden, *wieviel* Information der gestörte Ausgang einer *VVM* über den ungestörten Ausgang derselben *VVM* enthält. Als informationstheoretische Grundlage dient dabei zunächst ein sehr einfaches mathematisches Modell, so dass auf Standardlösungen zurückgriffen werden kann.

8.1.1 Das Shannonsche Kanalmodell und die Transinformation

Das Shannonsche Kanalmodell berücksichtigt 2 verschiedene Arten von Störungen. Prinzipiell können beide Arten gleichzeitig auftreten. Zunächst soll aber durch die getrennte Betrachtung eine Interpretation der verschiedenen Störungen gefunden werden.

Zum einen können verschiedene Eingangssymbole $X = x_i$ auf ein Ausgangssymbol $Y = y_j$ abgebildet werden, dies nennt man Äquivokation $H(X|Y)$. Diese Art von Störung kann als verlustbehaftete Signalkodierung verstanden werden (z.B. Quantisierung). Ist diese Störung gleich Null, so entspricht jedem Eingangssymbol genau ein zugehöriges Ausgangssymbol d.h. die Codierung ist verlustfrei. Im allgemeinen Fall gilt jedoch für die übertragene Information:

$$I(X,Y) = H(X) - H(X \mid Y) \tag{8.1.1}$$

Die Differenz $H(X) - H(X|Y)$ misst die durch einen Kanal übertragene Information über das Sendesymbol X bei empfangenem Empfangssymbol Y und wird als Transinformation *(engl. **M**utual **I**nformation)* bezeichnet (vergl. Anhang A).

Zum anderen kann ein Sendesymbol $X = x_i$ auf verschiedene Empfangssymbole $Y = y_j$ abgebildet werden, dies führt auf die Irrelevanz *H(Y|X)*. Störungen dieser Art können nur durch eine von außen auf das Eingangssymbol einwirkende Störung erklärt werden.

$$I(X,Y) = H(Y) - H(Y \mid X) \qquad (8.1.2)$$

In beiden Fällen lassen sich die bedingten Entropien *H(X|Y)* und *H(Y|X)* als Maß für Störungen interpretieren. Wird die Entropiedifferenz (8.1.1) auf *H(X)* normiert, kann im Prinzip der Verlust an Information angegeben werden, der bei der Rekonstruktion des Signals aus dem kodierten Signal entsteht. Es liegt dann ein Maß für die Effizienz einer Kodierung vor. Normiert man (8.1.2) auf *H(Y),* so gibt dieser Ausdruck den Verlust an Information an, der bei Störungen von außen entsteht. Dieses Maß quantifiziert also die Robustheit einer Kodierung gegenüber äußeren Störungen.

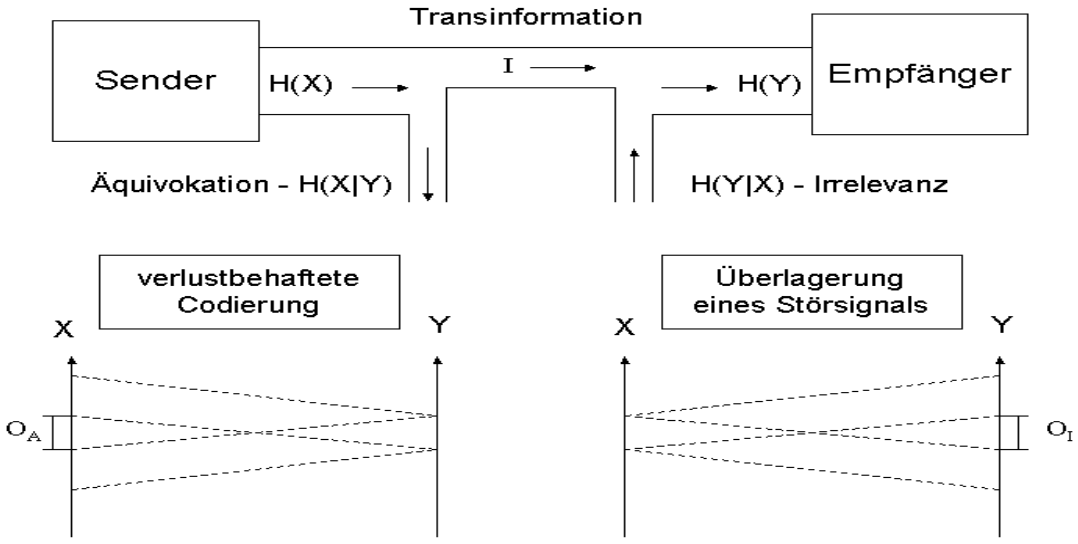

Abbildung 8.1.1 ***Mit dem Begriff der Äquivokation wird der Sachverhalt erfasst, dass eine Untermenge der Sendesymbole X auf ein Empfangssymbol $Y=y_j$ abgebildet wird. Die Irrelevanz dagegen hat zur Folge, dass ein Sendesymbol $X= x_i$ auf eine Untermenge der Empfangssymbole Y abgebildet wird.***

8.1.2 Informationstheoretisch motiviertes Robustheitsmaß

Bei Störungen sind die Empfangssymbole *Y* vollständig bestimmt, wenn die Sendesymbole *X* und die Störung *N* bekannt sind. Sind *X* und *N* darüber hinaus auch noch statistisch unabhängig voneinander, dann gilt für die Verbundentropie:

$$H(X,Y) = H(X) + H(N) \qquad (8.1.3.)$$

Der Zusammenhang zwischen der Verbundentropie und der Transinformation ist nach *[Cover-91]* durch die Gleichung 8.1.4 gegeben:

$$I(X,Y) = H(X) + H(Y) - H(X,Y) \tag{8.1.4}$$

Setzt man (8.1.3) in Gleichung (8.1.4) ein, so erhält man die folgende Beziehung:

$$I(X,Y) = H(Y) - H(N) \tag{8.1.5}$$

Es ist sinnvoll eine Normierung zu verwenden, mit der die Transinformation innerhalb fester Grenzen definiert werden kann, man spricht dann von einer relativen Transinformation.

$$I_{rel}(X,Y) = \frac{H(Y) - H(Y \mid X)}{H(Y)} = \frac{H(Y) - H(N)}{H(Y)} \tag{8.1.6}$$

Wie oben bereits erwähnt, stellt die bedingte Entropie *H(Y|X)* ein Maß für die Störung des Kanals dar. Im störungsfreien Fall verschwindet dieser Term und die relative Transinformation nimmt den Wert Eins an. Im anderen Fall *H(Y|X)=H(Y)*, d.h. bei Störungen die zur völligen Unabhängigkeit von *X* und *Y* führen, verschwindet die relative Transinformation und nimmt den Wert Null an.

Im Allgemeinen berücksichtigt die Transinformation auch nichtlineare Beziehungen zwischen zwei Zufallsvariablen. Da sowohl die auditive Signalverarbeitungskette als auch die Methoden zur Störreduktion nichtlineare Komponenten enthalten, scheint mit der *MI*-Analyse prinzipiell ein geeignetes Handwerkszeug zur Verfügung zu stehen, mit dem die Robustheit verschiedener *VVM* untersucht werden kann.

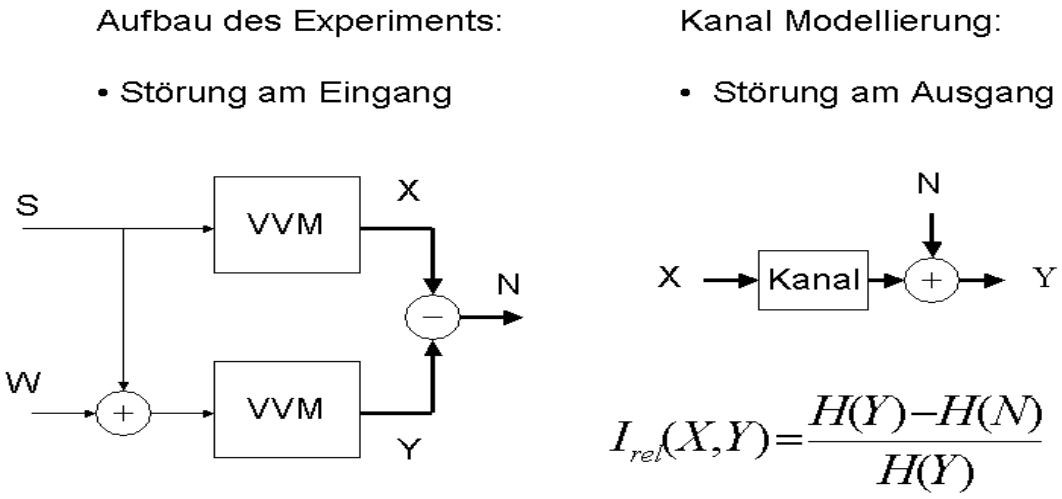

Abbildung 8.1.2 ***Der linke Teil zeigt den Aufbau des Experiments, der rechte Teil zeigt das zugrunde liegende Modell. Mit dem informationstheoretischen Modell kann die Ähnlichkeit der Größen X und Y in Abhängigkeit von der Stärke der Störsignals W am Eingang beurteilt werden.***

Im nächsten Abschnitt wird der vorgestellte informationstheoretische Ansatz auf gestörte Teilbandsignale erweitert. Die Auswirkungen einer äußeren Störung *W* auf das Sprachsignal *S* werden nun erst auf der Ebene der Teilbandsignale untersucht. Diese Modellierung entspricht einem parallelen Kanalmodell mit mehrdimensionalen Ein- und Ausgangsgrößen sowie einer nunmehr in die *VVM* verlagerte mehrdimensionale Störung ***N***. Den Aufbau eines Experiments zur Bestimmung der relativen Transinformation zeigt Abbildung 8.1.2.

8.1.3 RMI eines einzelnen Gaußschen Kanals

Geht man zunächst sowohl für das Signal *X* als auch für die Störung *N* von gaussverteilten Signalen aus, so erhält man für die Entropie der Ausgangsgröße *Y*:

$$H(Y) = \frac{1}{2} ld(2\pi e \sigma_Y^2) \tag{8.1.7}$$

Die Transinformation wird nach *[Cover-91]* durch den folgenden Ausdruck bestimmt:

$$I(X,Y) = \frac{1}{2} ld(1+SNR) \qquad \text{, mit} \quad SNR = \frac{\| X \|_2^2}{\| X - Y \|_2^2} \tag{8.1.8}$$

Somit erhält man nun für die relative Transinformation eines einzelnen Gaußschen Kanals:

$$I_{rel}(X,Y) = \frac{ld(1+SNR)}{ld(2\pi e \sigma_Y^2)} \tag{8.1.9}$$

8.1.4 RMI von parallelen unabhängigen Gaußschen Kanälen

In Multibandmodellen muss mit multivariaten Zufallsgrößen gerechnet werden, dann gilt nach *[Cover-91]* die Beziehung (8.1.10):

$$H(\mathbf{Y}) = H(Y_1, Y_2, ..., Y_N) \le \sum H(Y_i) \tag{8.1.10}$$

Das Gleichheitszeichen gilt bei statistischer Unabhängigkeit der Teilbänder, dann liegt die Kovarianzmatrix **C** einer gaussverteilten multivariaten Zufallsgröße als Diagonalmatrix vor, deren Diagonalelemente $C_{ii} = P_Y^i$ den Leistungen in den einzelnen Teilbändern entsprechen. Im allgemeinen Fall gilt jedoch für die Entropie der multivariaten Ausgangsgröße *Y* die folgende Gleichung:

$$H(\mathbf{Y}) = \frac{1}{2} ld((2\pi e)^n |\mathbf{C}|) \tag{8.1.11}$$

Durch die Anwendung eines *LIN* bzw. der *DCT* werden voneinander unabhängige Teilbänder angestrebt. Geht man von dieser Approximation aus, so führt $|\mathbf{C}| = \prod_N C_{ii}$ zu einer Vereinfachung:

$$H(\mathbf{Y}) = \frac{1}{2}\left[ld(2\pi e)^N + ld(|\mathbf{C}|)\right] = \frac{1}{2}[ld(2\pi e)^N + \sum_{i=1}^{N} ld(P_Y^i)] \tag{8.1.12}$$

Für die Transinformation von parallelen unabhängigen Gaußschen Kanälen gilt nach *[Cover-91]* ähnlich wie im eindimensionalen Fall:

$$I(\mathbf{X},\mathbf{Y}) = \frac{1}{2}\sum_{i=1}^{N} ld(1+SNR^i) \tag{8.1.13}$$

Das SNR in den einzelnen Teilbändern erhält man mit (8.1.14):

$$SNR^i = \frac{\| X^i \|_2^2}{\| X^i - Y^i \|_2^2} = \frac{P_X^i}{P_N^i} \tag{8.1.14}$$

Somit kann für die relative Transinformation von multivariaten zeitabhängigen Signalen der folgende Ausdruck angegeben werden:

$$I_{rel}(X,Y) = \frac{\sum_{i=1}^{N} ld(1+SNR^i)}{ld(2\pi e)^N + \sum_{i=1}^{N} ld(P_Y^i)} \tag{8.1.15}$$

Dieses Robustheitsmaß hängt nur von den Teilband SNR und den Teilbandleistungen ab. Allerdings sind hier theoretische Annahmen über die Verteilung der Zufallsgrößen gemacht worden, die praktisch nicht eingehalten werden können.

8.2 Definition der Robustheit gemäß der Artikulationstheorie

8.2.1 Die Phonartikulation in den Teilbändern

Für das Multibandmodell konnte die Phonartikulation *s* mit der folgenden Gleichung berechnet werden (Anhang B).

$$s = 1 - \prod_{i=1}^{N} e_i \tag{8.2.1}$$

Dabei wird mit e_i der Artikulationsfehler im *i*-ten Teilband bezeichnet. Der Artikulationsfehler in den Teilbändern und damit auch die Phonartikulation kann durch Änderung des eingangsseitigen SNR über den Verstärkungsfaktor $\alpha \in [0...1]$ von *α* abhängig gemacht werden. Zur Berechnung der Artikulationsfehler in den einzelnen Teilbändern gibt die Artikulationstheorie mit $e_{min} \approx 0.015$ den folgenden Zusammenhang an:

$$e_i = e^{D_i \min}, \quad \text{mit} \quad D_i(\alpha) = \frac{1}{N} SNR_i(\alpha)/30\,dB \tag{8.2.2}$$

Demnach hängt D_i als Funktion von *α* lediglich vom SNR im i-ten Teilband ab. Außerdem werden die Teilband SNR auf 30db bzw. 0dB begrenzt, so dass man bei α → 1 für den Gesamtartikulationsfehler den Wert $e = e_{\min}$ bzw. für die Phonartikulation den Wert $s_{\max} = 1 - e_{\min}$ erhält. Im entgegen gesetzten Fall α → 0 nehmen alle Teilband SNR den Wert Null an, so dass die Phonartikulation ebenfalls den Wert Null annimmt.

8.2.2 Analogie zur Zuverlässigkeit von Parallelsystemen

Die mathematische Struktur von (8.2.1) ist anlog zur Beschreibung der Zuverlässigkeit von Parallelsystemen mit *N* Komponenten, welche unabhängig voneinander ausfallen können. Nach *[Weinrichter-91]* wird mit *R* die Zuverlässigkeit des Gesamtsystems, mit $\underline{p_i}$ die Zuverlässigkeit der Komponente *i* und mit q_i – die Ausfallwahrscheinlichkeit der Komponente *i* bezeichnet.

$$R = 1 - \prod_{i=1}^{N} (1 - p_i) \qquad bzw. \qquad R = 1 - \prod_{i=1}^{N} q_i \tag{8.2.3}$$

Die Ausfallwahrscheinlichkeit q_i einer Parallelkomponente kann im Allgemeinen folgende Werte annehmen:

$$q_i \in \{0...1\} \;\Rightarrow\; R \in \{1...0\} \tag{8.2.4}$$

Dagegen kann die Fehlerwahrscheinlichkeit e_i eines Artikulationsbandes *i* nur die folgenden Werte annehmen:

$$D_i \in \{0...1/N\} \;\Rightarrow\; e_i \in \{1...e_{\min}^{1/N}\} \tag{8.2.5}$$

Die oben angesprochene Analogie rechtfertigt den Ansatz, zur Beurteilung der Robustheit einer *VVM* die durch das Multibandmodell definierte Zuverlässigkeit zu verwenden. Bei dieser Definition wird lediglich davon ausgegangen, dass die Signale in den Teilbändern unabhängig voneinander verarbeitet werden, darüber hinaus werden keine weiteren Annahmen über die statistischen Verteilungen der Teilbandsignale gemacht.

8.3 Beschreibung des Vorexperiments

In dem Vorexperiment wird jedes Audiosignal mit verschiedenen Störgeräuschen unterschiedlicher Stärke überlagert und die Zuverlässigkeit ermittelt. Anschließend wird für die jeweilige *VAMIG*- Ausprägung der Mittelwert der Zuverlässigkeit ermittelt. Diese statistische Größe, wird nun für jede Störgeräuschsituation gegen das SNR am Eingang aufgetragen. Das Experiment wird anschließend für verschiedene *VAMIG*- Ausprägungen durchgeführt, so dass die verschiedenen Ausprägungen miteinander verglichen werden können. Zur Durchführung des Experiments wird auf die Testmenge A der Datenbasis Aurora2 zurückgegriffen. Diese Testmenge enthält etwa 1000 ungestörte Audiodateien für jeweils 4 unterschiedliche Geräuschkategorien (SUBWAY, CAR, BABBLE, EXHIBITION). Die Störungen wurden den Audiodateien der ungestörten Testmenge mit verschiedenen SNR im Bereich [20, 15, 10, 5, 0, -5] dB zugemischt. Zusätzlich zum Aurora Umfang wurde für jedes Testset eine fünfte Störung erzeugt, dabei handelt es sich um weißes Rauschen.

Die folgenden Tabellen zeigen die gemessenen Phonartikulationen für die Ausprägungen *HC-LIN*, *HC-DCT* und *LOG-DCT* jeweils für die Filterbankvarianten PERDIGAO (Frequenzbereich, Gammatone), SHAMMA (Zeitbereich, asymmetrische Filter) und Wavelet-Zerlegung (Zeitbereich). Die Anzahl der Filterbankkanäle wurde für alle Varianten auf 21 festgesetzt. Für die ersten beiden Filterbankvarianten sind die Ergebnisse sowohl für inaktive bzw. aktive Bindung (*PKF*-Mode) angegeben. Leider wurde für die Wavelet-Zerlegung kein befriedigend arbeitender Algorithmus zur Erzeugung von pitchkohärenten Merkmalen gefunden. Hier wird daher zwischen konventionellen Energiemerkmalen und den *TEO*-Merkmalen unterschieden. Für alle Ausprägungen erfolgen die Untersuchungen zunächst ohne Spektrale Subtraktion. In diesem Fall kann die Robustheit der Merkmale tatsächlich auf die auditorischen Eigenschaften der jeweiligen *VAMIG*-Ausprägung und nicht auf die Geräuschreduktion zurückgeführt werden.

Geräuschkategorie SUBWAY

	20 dB	15 dB	10 dB	5 dB	0 dB	-5dB
PERDIGAO	0.981/ 0.909	0.972/ 0.895	0.956/ 0.875	0.934/ 0.849	0.906/ 0.820	0.876/ 0.786
SHAMMA	0.976/ 0.948	0.968/ 0.938	0.952/ 0.923	0.923/ 0.901	0.907/ 0.874	0.880/ 0.843
WAVELET	0.830/ 0.828	0.807/ 0.804	0.779/ 0.775	0.745/ 0.739	0.708/ 0.702	0.665/ 0.658

Tabelle 8.3.1: ***VAMIG-Ausprägung HC-LIN; PKF-Mode off/on; TEO-Mode off/on;***

	20 dB	15 dB	10 dB	5 dB	0 dB	-5dB
PERDIGAO	0.941/ 0.814	0.905/ 0.787	0.885/ 0.749	0.795/ 0.704	0.727/ 0.651	0.661/ 0.596
SHAMMA	0.936/ 0.812	0.900/ 0.787	0.848/ 0.747	0.787/ 0.693	0.714/ 0.632	0.641/ 0.596
WAVELET	0.951/ 0.935	0.922/ 0.900	0.881/ 0.851	0.832/ 0.794	0.774/ 0.727	0.717/ 0.664

Tabelle 8.3.2: ***VAMIG-Ausprägung HC-DCT; PKF-Mode off/on; TEO-Mode off/on;***

	20 dB	15 dB	10 dB	5 dB	0 dB	-5dB
PERDIGAO	0.870/ 0.841	0.857/ 0.794	0.844/ 0.734	0.844/ 0.664	0.827/ 0.586	0.810/ 0.503
SHAMMA	0.863/ 0.828	0.854/ 0.780	0.843/ 0.714	0.829/ 0.632	0.813/ 0.540	0.792/ 0.445
WAVELET	0.885/ 0.871	0.870/ 0.856	0.856/ 0.842	0.841/ 0.827	0.826/ 0.813	0.812/ 0.800

Tabelle 8.3.3 ***VAMIG-Ausprägung LOG-DCT; PKF-Mode off/on; TEO-Mode off/on;***

Geräuschkategorie BABBLE

	20 dB	15 dB	10 dB	5 dB	0 dB	-5dB
PERDIGAO	0.982 / 0.906	0.975 / 0.891	0.962 / 0.870	0.942 / 0.842	0.917 / 0.810	0.888 / 0.773
SHAMMA	0.977 / 0.945	0.971 / 0.935	0.958 / 0.919	0.939 / 0.895	0.915 / 0.866	0.888 / 0.834
WAVELET	0.824 / 0.823	0.802 / 0.799	0.773 / 0.769	0.739 / 0.734	0.698 / 0.692	0.652 / 0.646

Tabelle 8.3.4: ***VAMIG-Ausprägung HC-LIN; PKF-Mode off/on; TEO-Mode off/on;***

	20 dB	15 dB	10 dB	5 dB	0 dB	-5dB
PERDIGAO	0.934 / 0.806	0.897 / 0.775	0.847 / 0.733	0.783 / 0.680	0.714 / 0.626	0.644 / 0.568
SHAMMA	0.928 / 0.802	0.889 / 0.772	0.838 / 0.726	0.773 / 0.667	0.701 / 0.601	0.628 / 0.535
WAVELET	0.947 / 0.930	0.916 / 0.892	0.874 / 0.841	0.819 / 0.779	0.758 / 0.712	0.697 / 0.645

Tabelle 8.3.5: ***VAMIG-Ausprägung HC-DCT; PKF-Mode off/on; TEO-Mode off/on;***

	20 dB	15 dB	10 dB	5 dB	0 dB	-5dB
PERDIGAO	0.863 / 0.827	0.851 / 0.776	0.836 / 0.711	0.818 / 0.635	0.800 / 0.552	0.781 / 0.465
SHAMMA	0.861 / 0.810	0.852 / 0.754	0.842 / 0.682	0.826 / 0.595	0.808 / 0.496	0.786 / 0.401
WAVELET	0.881 / 0.866	0.867 / 0.853	0.853 / 0.839	0.838 / 0.825	0.822 / 0.811	0.809 / 0.799

Tabelle 8.3.6: ***VAMIG-Ausprägung LOG-DCT; PKF-Mode off/on; TEO-Mode off/on;***

Geräuschkategorie CAR

	20 dB	15 dB	10 dB	5 dB	0 dB	-5dB
PERDIGAO	0.980 / 0.911	0.972 / 0.895	0.955 / 0.875	0.932 / 0.846	0.905 / 0.814	0.874 / 0.778
SHAMMA	0.977 / 0.948	0.969 / 0.938	0.953 / 0.922	0.932 / 0.897	0.905 / 0.866	0.874 / 0.833
WAVELET	0.822 / 0.820	0.796 / 0.793	0.768 / 0.764	0.729 / 0.724	0.688 / 0.683	0.643 / 0.637

Tabelle 8.3.7: ***VAMIG-Ausprägung HC-LIN; PKF-Mode off/on; TEO-Mode off/on;***

	20 dB	15 dB	10 dB	5 dB	0 dB	-5dB
PERDIGAO	0.937 / 0.811	0.901 / 0.779	0.851 / 0.737	0.790 / 0.686	0.722 / 0.629	0.656 / 0.575
SHAMMA	0.935 / 0.804	0.898 / 0.772	0.849 / 0.727	0.788 / 0.666	0.716 / 0.593	0.646 / 0.527
WAVELET	0.950 / 0.932	0.920 / 0.893	0.877 / 0.843	0.824 /0.781	0.767 / 0.716	0.710 / 0.652

Tabelle 8.3.8: ***VAMIG-Ausprägung HC-DCT; PKF-Mode off/on; TEO-Mode off/on;***

	20 dB	15 dB	10 dB	5 dB	0 dB	-5dB
PERDIGAO	0.871 / 0.827	0.858 / 0.775	0.844 / 0.707	0.827 / 0.631	0.809 / 0.546	0.791 / 0.462
SHAMMA	0.869 / 0.811	0.860 / 0.754	0.850 / 0.682	0.835 / 0.589	0.817 / 0.483	0.795 / 0.388
WAVELET	0.884 / 0.871	0.871 / 0.856	0.857 / 0.843	0.842 / 0.828	0.827 / 0.814	0.815 / 0.803

Tabelle 8.3.9: ***VAMIG-Ausprägung LOG-DCT; PKF-Mode off/on; TEO-Mode off/on;***

Geräuschkategorie EXHIBITION

	20 dB	15 dB	10 dB	5 dB	0 dB	-5dB
PERDIGAO	0.980/ 0.906	0.971/ 0.890	0.955/ 0.869	0.933/ 0.841	0.906/ 0.810	0.876/ 0.775
SHAMMA	0.973/ 0.942	0.963/ 0.930	0.947/ 0.913	0.926/ 0.887	0.901/ 0.861	0.873/ 0.775
WAVELET	0.837/ 0.836	0.813/ 0.811	0.788/ 0.784	0.755/ 0.750	0.715/ 0.709	0.675/ 0.669

Tabelle 8.3.10: ***VAMIG-Ausprägung HC-LIN; PKF-Mode off/on; TEO-Mode off/on;***

	20 dB	15 dB	10 dB	5 dB	0 dB	-5dB
PERDIGAO	0.935/ 0.809	0.898/ 0.780	0.848/ 0.743	0.786/ 0.691	0.720/ 0.737	0.652/ 0.577
SHAMMA	0.931/ 0.801	0.893/ 0.770	0.841/ 0.730	0.775/ 0.669	0.704/ 0.603	0.628/ 0.531
WAVELET	0.949/ 0.934	0.918/ 0.897	0.877/ 0.849	0.825/ 0.788	0.769/ 0.725	0.710/ 0.659

Tabelle 8.3.11: ***VAMIG-Ausprägung HC-DCT; PKF-Mode off/on; TEO-Mode off/on;***

	20 dB	15 dB	10 dB	5 dB	0 dB	-5dB
PERDIGAO	0.864/ 0.836	0.851/ 0.785	0.838/ 0.726	0.821/ 0.647	0.806/ 0.566	0.788/ 0.478
SHAMMA	0.857/ 0.820	0.848/ 0.765	0.837/ 0.700	0.823/ 0.605	0.807/ 0.508	0.787/ 0.404
WAVELET	0.887/ 0.873	0.872/ 0.858	0.858/ 0.844	0.842/ 0.829	0.828/ 0.814	0.812/ 0.800

Tabelle 8.3.12: ***VAMIG-Ausprägung LOG-DCT; PKF-Mode off/on; TEO-Mode off/on;***

Geräuschkategorie WHITE NOISE

	20 dB	15 dB	10 dB	5 dB	0 dB	-5dB
PERDIGAO	0.952 / 0.888	0.933 / 0.871	0.908 / 0.848	0.878 / 0.821	0.847 / 0.791	0.816 / 0.758
SHAMMA	0.946 / 0.931	0.929 / 0.919	0.907 / 0.903	0.880 / 0.881	0.852 / 0.855	0.816 / 0.826
WAVELET	0.791 / 0.787	0.760 / 0.754	0.723 / 0.716	0.681 / 0.673	0.632 / 0.624	0.595 / 0.587

Tabelle 8.3.13: ***VAMIG-Ausprägung HC-LIN; PKF-Mode off/on; TEO-Mode off/on;***

	20 dB	15 dB	10 dB	5 dB	0 dB	-5dB
PERDIGAO	0.881 / 0.786	0.838 / 0.754	0.786 / 0.711	0.728 / 0.662	0.667 / 0.607	0.607 / 0.551
SHAMMA	0.882 / 0.782	0.836 / 0.749	0.783 / 0.702	0.722 / 0.646	0.667 / 0.579	0.591 / 0.513
WAVELET	0.888 / 0.863	0.853 / 0.816	0.808 / 0.757	0.757 / 0.696	0.705 / 0.633	0.654 / 0.576

Tabelle 8.3.14: ***VAMIG-Ausprägung HC-DCT; PKF-Mode off/on; TEO-Mode off/on;***

	20 dB	15 dB	10 dB	5 dB	0 dB	-5dB
PERDIGAO	0.866 / 0.778	0.854 / 0.725	0.841 / 0.662	0.828 / 0.588	0.814 / 0.507	0.798 / 0.421
SHAMMA	0.868 / 0.765	0.859 / 0.707	0.849 / 0.635	0.838 / 0.550	0.824 / 0.451	0.805 / 0.354
WAVELET	0.875 / 0.858	0.861 / 0.845	0.846 / 0.830	0.832 / 0.817	0.818 / 0.802	0.804 / 0.789

Tabelle 8.3.15: ***VAMIG-Ausprägung LOG-DCT; PKF-Mode off/on; TEO-Mode off/on;***

Mittlere Phonartikulation

Um eine bessere Übersicht zur Bewertung der *VAMIG*-Ausprägungen zu erhalten, zeigen die folgenden drei Tabellen nun die Mittelwerte über alle Geräuschkategorien.

	20 dB	15 dB	10 dB	5 dB	0 dB	-5dB
PERDIGAO	0.975 / 0.904	0.964 /0.888	0.947 / 0.867	0.924 / 0.840	0.896 / 0.809	0.866 /0.774
SHAMMA	0.970 / 0.943	0.960 / 0.923	0.943 / 0.916	0.922 / 0.829	0.896 / 0.864	0.869 / 0.833
WAVELET	0.821 / 0.819	0.796 / 0.792	0.766 / 0.762	0.729 / 0.724	0.688 / 0.682	0.646 / 0.639

Tabelle 8.3.16: ***VAMIG-Ausprägung HC-LIN; PKF-Mode off/on; TEO-Mode off/on;***

	20 dB	15 dB	10 dB	5 dB	0 dB	-5dB
PERDIGAO	0.925 / 0.805	0.888 / 0.775	0.837 / 0.735	0.776 / 0.685	0.710 / 0.630	0.644 / 0.573
SHAMMA	0.922 / 0.800	0.883 / 0.770	0.837 / 0.727	0.776 / 0.668	0.710 / 0.601	0.644 / 0.534
WAVELET	0.937 / 0.919	0.906 / 0.879	0.864 / 0.825	0.811 / 0.768	0.755 / 0.702	0.697 / 0.639

Tabelle 8.3.17: VAMIG-Ausprägung HC-DCT; PKF-Mode off/on; TEO-Mode off/on;

	20 dB	15 dB	10 dB	5 dB	0 dB	-5dB
PERDIGAO	0.867 / 0.822	0.854 / 0.771	0.841 / 0.708	0.825 / 0.633	0.808 / 0.551	0.790 / 0.468
SHAMMA	0.864 / 0.807	0.855 / 0.752	0.844 / 0.683	0.830 / 0.594	0.814 / 0.496	0.793 / 0.398
WAVELET	0.882 / 0.868	0.868 / 0.854	0.854 / 0.839	0.839 / 0.825	0.824 / 0.811	0.810 / 0.798

Tabelle 8.3.18: VAMIG-Ausprägung LOG-DCT; PKF-Mode off/on; TEO-Mode off/on;

8.4 Auswertung des Vorexperiments und Schlussfolgerungen

Bei ausgeschalteter Spektraler Subtraktion schneiden die rein auditiven Methoden (*HC-LIN*) – mit Ausnahme der Wavelet Realisierung – für alle Störungen deutlich besser ab als die Vergleichsausprägungen. Es hat also zunächst den Anschein, dass man im Gegensatz zur klassischen *LOG-DCT* Ausprägung für die *HC-LIN* Ausprägung eine engere Bindung zur menschlichen Signalverarbeitung beobachtet kann. Die gemessene Zuverlässigkeit für pitchkohärente Merkmale blieb leider deutlich hinter den Erwartungen zurück. Allerdings ist zu berücksichtigen, dass die Berechnung von pitchkohärenten Merkmalen zunächst nur nach dem Prinzip der primitiven Bindung erfolgte *[Römer-06]*. Eine spätere Analyse der Modellannahmen zeigte, dass das Modell der primitiven Bindung einige gravierende Nachteile aufweist.

Bei der Berechnung der Teilbandkorrelationen galt nämlich die Annahme, dass diese in rauschenden Segmenten mit zunehmendem Verschiebungsindex schnell abfällt. Diese Annahme kann aber nur für breitbandige Filter aufrechterhalten werden. Die Übertragungsfunktionen der Gammatone-Filterbank sind insbesondere für die tiefen Frequenzen schmalbandig, so dass in diesen Bändern nur impulsartige Teilbandspektren aufgelöst werden können. Impulsartige Leistungsspektren werden durch die *IFFT* aber in kosinusförmige Autokorrelationsfunktionen transformiert, die an der Stelle $r_i[M]$ sehr häufig ungleich Null sind. D.h. selbst wenn sich in einem Teilband nur Rauschenergie befindet, so kann der normierte *AKF*-Wert an der Stelle M durchaus Maximalwerte erreichen. Dies hat nun zur Folge, dass die Gewichtungsfaktoren in den stimmlosen Segmenten bei der primitiven Bedingung als multiplikative Rauschsignale wirksam werden und dem erwarteten Gewinn in den stimmhaften Abschnitten entgegenwirken. Die Idee breitbandigere Filter zu verwenden, erscheint nur auf den ersten Blick sinnvoll. Die zunehmende Überlappung der Filter führt dann dazu, dass sich die Harmonischen der Sprachgrundfrequenz nicht mehr auflösen lassen. In Folge dessen verlieren daher auch die Gewichtungsfaktoren ihre Information. Für die später durchgeführten Erkennexperimente auf Ziffern wurde daher ausschließlich das Modell der bedingten Bindung verwendet.

Die bei der Wavelet-Zerlegung beobachtete schlechte Schätzqualität der dominanten Frequenz (verursacht durch die hohe Unterabtastung in einigen Bändern) führte bei der *LIN*-Ausprägung des Dekorrelations-Moduls zu einem Abfall der Zuverlässigkeit. Dieser Abfall lässt sich auf einen Verstoß gegen das Prinzip des „Continuity Preserving Signal Processing“ *[Andringa-02]* zurückführen. Dagegen konnte die Wavelet-Zerlegung bei den hybriden *VAMIG*- Ausprägungen (*HC-DCT*) die besten Resultate erzielen. Da für die Wavelet-Filterbank aber noch kein befriedigend arbeitender Algorithmus zur Erzeugung von pitchkohärenten Merkmalen vorlag und auch mit den *TEO*-Merkmalen kein signifikanter Zuverlässigkeitsgewinn erzielt werden konnte, wurde dieser Ansatz nicht weiter verfolgt.

Der erfolgversprechendste Ansatz basiert nach den hier vorliegenden Experimenten auf der *HC-LIN* Ausprägung auf der Grundlage einer Gammatone-Filterbankanalyse. Da diese *VAMIG*-Ausprägung im Frequenzbereich arbeitet, können nun auch die hochauflösenden Methoden zur Realisierung der Bindungseigenschaft untersucht werden.

9 Untersuchungen zur Erkenngenauigkeit

In diesem Kapitel erfolgt zunächst eine kurze Einführung in die *HMM*-Sprachtechnologie. Dabei werden zum besseren Verständnis nur zwei elementare Anwendungen beschrieben. Die Erkennung von isoliert gesprochenen Worten und die Verbundworterkennung. Insbesondere für die Untersuchungen zur Erkenngenauigkeit der verschiedenen *VAMIG*-Ausprägungen sollte nicht auf Kontextwissen bei der Verbundworterkennung zurückgegriffen werden. Daher werden die in Abschnitt 9.2 beschriebenen Experimente ebenfalls mit Ziffernketten der Aurora-2 Datenbasis durchgeführt.

9.1 Modellierung und Erkennung mit der HMM-Technologie

Die meisten Spracherkennungssysteme basieren derzeit auf der flexiblen ***H**idden **M**arkov **M**odel*-Technologie *[Fink-03]*. Dabei geht man zum einen von der Annahme aus, dass zur Erzeugung von lautsprachlichen Einheiten nicht immer eine feste Anzahl von Zuständen durchlaufen werden muss. Zum anderen sollte die Erzeugung von Lauten in einem Zustand gemäß einer Wahrscheinlichkeitsverteilung erfolgen. Diese beiden Annahmen, mit denen letztlich die Vielfalt sprachlicher Äußerungen berücksichtigt werden soll, erfordern flexible Modellstrukturen. Für den Aufbau und die Benutzung solcher Modelle müssen zunächst zwei Phasen unterschieden werden: Die Erkennphase und eine Trainingsphase. In der Trainingsphase werden die Parameter der *HMM*-Modelle λ^i erlernt.

- Anzahl der Zustände: N
- Matrix von Zustandsübergängen: $\mathbf{A} = \{a^i_{s,s'}\}$
- Matrix von Emissionsverteilungen: $\mathbf{B} = \{b^i_s(\mathbf{x})\}$ (9.1)
- Initiale Zustandsverteilung: $\Pi = \{\pi^i_s\}$
- Mixture Gewichte: $c^i_{s,k}$

Die Modellierung der Emmisionsverteilung eines Zustands *s* beruht häufig auf Gaußschen gewichteten Mixturen. D.h. jeder Zustand eines Modells λ^i wird durch eine Mixtur von *K*-multivariaten Gaußverteilungen beschrieben:

$$b^i_s = p(\mathbf{x} \mid i, s) = \sum_K c^i_{s,k} N(\mathbf{x}, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{9.2}$$

Dabei fordert das Axiom 2 (Anhang A), dass die Bedingung (9.3) eingehalten werden muss.

$$\sum_k c^i_{s,k} = 1 \tag{9.3}$$

Häufig werden die Parameter $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$ für die Modelle λ^i zugunsten der Speichereffizienz nicht mehr unterschieden sondern in einem gemeinsamen Codebuch abgelegt. Die verschiedenen Ausprägungen der Emissionsverteilungen werden dann lediglich über unterschiedliche Mixture Gewichte realisiert.

In der Erkennphase muss zunächst das zu erkennende Vokabular definiert werden. Anschließend wird das Vokabular in eine Folge von *HMM*-Modellen übersetzt. Im einfachsten Fall werden einfache Schleifen ohne Verwendung von Sprachmodellen gebildet (Ziffern-Loop, Phonem-Loop), die dann den Suchraum aufspannen.

Der Erkennprozess kann nun wie folgt beschrieben werden: Mit jedem Frame liefert die *VVM* einen Merkmalsvektor ***x***. Unter Verwendung der ***B***-Matrix kann nun für jedes Wort des Vokabulars in dessen aktuellem Zustand die Emissionswahrscheinlichkeit berechnet werden. Anschließend erfolgt der Übergang zum nächsten Zustand gemäß der ***A***-Matrix. Dieser Prozess kann unter Zuhilfenahme einer sogenannten Trellis beschrieben werden. Mit dem Viterbi-Algoritmus kann nun für jedes Wort des Vokabulars der optimale Pfad durch die Trellis berechnet werden. Möchte man nur isoliert gesprochene Wörter erkennen, so wird der Suchraum einfach durch parallele Anordnung der *HMM*-Folgen aufgespannt. Das erkannte Wort ist daher das Wort mit der höchsten Wahrscheinlichkeit.

Soll eine ganze Kette von Worten erkannt werden, so kann der Suchraum im einfachsten Fall unter Verwendung einer Schleife von *HMM*-Folgen aufgebaut werden. Hier wird dann die wahrscheinlichste Wortkette ausgewählt. Sollen grammatikalisch korrekte Wortfolgen erkannt werden, so können grammatikalische Regeln beim Aufbau des Suchraums berücksichtigt werden (*Command and Control*). Können diese Regeln nicht vorgegeben werden, ist auch der Einsatz von statistischen Sprachmodellen möglich (*Dictation*).

Die Verbundworterkennung ist natürlich der praktisch bedeutsamere Anwendungsfall, insbesondere bei der Erkennung von langen Ziffernketten werden die menschlichen Erkennleistungen bereits übertroffen. Dies liegt natürlich daran, dass bei dieser Anwendung kein Kontextwissen angewendet werden kann. Genau diese Einschränkung ist aber sinnvoll bei der Untersuchung der Leistungsfähigkeit unterschiedlicher *VVM*. D.h. hier gilt die Forderung, dass solche Untersuchungen unabhängig vom individuellen Kontextwissen durchgeführt werden müssen. Eine weitere Anwendung ist bspw. die Aufnahme von dynamischem Vokabular, dies geschieht in der Regel durch das Abspeichern von Phonemketten *[Römer-02]*.

Ein allgemein akzeptiertes Gütekriterium für Spracherkennungstests ist die Wortakkuratheit, dabei werden die wortweisen Substitutionen, Insertionen und Deletionen berücksichtigt. Wenn also im Folgenden von Erkennraten die Rede ist, dann ist damit die Wortakkuratheit in Prozent gemeint:

$$WA = 1 - \frac{N_{sub} + N_{del} + N_{ins}}{N_{tot}} \tag{9.4}$$

Bei der darüber hinaus bekannten Wortkorrektheit geht die Anzahl der Insertionen nicht ein, diese Größe wird in den folgenden Experimenten nicht beachtet.

9.2 Erkennexperimente mit Ziffernketten

Alle in diesem Kapitel beschriebenen *VAMIG*- Ausprägungen transformieren das Sprachsignal (8 kHz) in Fenstern von 32 ms Länge in den Frequenzbereich. Die nachfolgenden Module arbeiten dann im Gegensatz zu den in *[Tchorz-99]* beschriebenen Experimenten lediglich mit einer Abtastrate von 125 Hz. Es erfolgt keine Preemphase und es werden keine speziellen Optimierungen zur Verbesserung der Erkennrate vorgenommen, so dass die Unterschiede in den Erkennraten tatsächlich auf den Austausch der Kernmodule des *VAMIG* zurückzuführen sind.

Als Referenz wird die Standard *MFCC*-Vorverarbeitung aus *[Voicebox]* verwendet. Diese Entscheidung muss kurz begründet werden. Die Vorverarbeitung von Aurora-2 ist bezüglich der Parametrierung kaum variierbar, dagegen lassen sich mit der oben genannten *MFCC*-Implementierung die Anzahl der Kanäle und die untere Grenzfrequenz der Mel-Filterbank frei einstellen. Dies ist einerseits eine notwendige Voraussetzung um die Vergleichbarkeit zwischen den Modellen zu gewährleisten. Andererseits muss für die Bindung gewährleistet sein, dass die Anzahl der Teilbänder genügend groß ist, um die Gewichtungsfaktoren mit einer genügend hohen Auflösung berechnen zu können.

Zunächst werden die Kernmodule der *MFCC*-Vorverarbeitung (39-Dreiecksfilterbank, statische Kompression) schrittweise durch die Module 39-Gammatone-Filterbank (200 Hz bzw. 75 Hz untere Grenzfrequenz) und dynamische Kompression ersetzt. Nach vollständiger Substitution und einem abschließendem Modulationstiefpass von 8 Hz Grenzfrequenz liegt ein rein auditives *PAS-Modell* vor. Die abschließende Dekorrelation der Merkmalsvektoren erfolgt einheitlich für alle *VAMIG*-Ausprägungen durch eine *DCT.* Die 13 *DCT* Koeffizienten werden um 13-Δ und 13-$\Delta\Delta$ Koeffizienten ergänzt und bilden den für das Training und die Erkennung verwendeten Supervektor. Die Δ und $\Delta\Delta$ Koeffizienten werden durch die 1. und 2. Differenz jeweils erster Ordnung bestimmt. Diese einfache Berechnung hat den Vorteil, dass keine zusätzliche Verzögerung in Kauf genommen werden muss. Die Konfiguration der *HMM*–Modelle entspricht *[Hirsch-00]*, nach einem Clean-Condition Training erfolgt die Erkennung ausschließlich auf dem Testset-A (BABBLE, CAR, SUBWAY, EXHIBITION). Die folgenden Abbildungen zeigen die mittleren Erkennraten der 5 Störszenarien ohne Schwellenanpassung, mit Schwellenanpassung und für Schwellenanpassung mit einer zusätzlichen Normalisierung der Supervektorkomponenten.

Die Schwellenanpassung für die Haarzellenmodelle erfolgt auf der *RMS*-Ebene gemäß der Vorschrift *OUT = MAX (INP-NL, A),* also noch vor der Kompressionsstufe. Für die statische Kompression (*LOG*) wurde *A=1.0* gesetzt, für die dynamische Kompression (Haarzelle) hat sich der Wert *A=3* bewährt. Mit der Variablen *NL* wird die Schätzung des *RMS*-Rauschniveaus im jeweiligen Teilband bezeichnet. Die Normalisierung der Komponenten des Supervektors erfolgt nach Vorschrift $OUT = (INP-\mu)/\delta$. Zunächst wurden neben dem *MFCC*-Standardmodell die Erkennraten der folgenden *VAMIG*-Ausprägungen ermittelt: *LOG-DCT* (Gammatone-Filterbank, statische Kompression, *DCT*), *HC-DCT* (Gammatone-Filterbank, dynamische Kompression, *DCT*) und *PAS* (Gammatone-Filterbank, dynamische Kompression, *AM*-Filter, *LIN*, *DCT*). Anschließend wurden die Erkennraten für die *PAS-CAS*-Ausprägungen ermittelt.

	INF	20 dB	15 dB	10 dB	5 dB	0 dB	-5dB
MFCC-SM0	97.79	76.54	63.60	46.14	24.59	10.53	6.75
LOG-DCT-SM0	96.94	82.07	73.29	55.62	27.17	11.28	8.61
HC-DCT-SM0	93.83	78.97	51.98	21.87	10.57	7.66	6.95
PAS-SM0	95.59	80.62	54.43	22.13	1.88.	0.61	3.38

Abbildung 9.1 ***Erkennraten der VAMIG Ausprägungen ohne Schwellenanpassung und ohne Normalisierung bei einer unteren Grenzfrequenz von 200 Hz***

	INF	20 dB	15 dB	10 dB	5 dB	0 dB	-5dB
MFCC-SM2	88.75	21.48	12.77	9.63	8.79	7.85	7.87
LOG-DCT-SM2	95.05	59.05	26.84	13.41	9.71	8.69	8.41
HC-DCT-SM2	93.59	87.61	73.95	45.13	17.93	7.89	4.84
PAS-SM2	95.20	91.78	83.89	65.51	35.91	12.44	6.21

Abbildung 9.2 ***Erkennraten der VAMIG Ausprägungen mit Schwellenanpassung und ohne Normalisierung bei einer unteren Grenzfrequenz von 200 Hz***

	INF	20 dB	15 dB	10 dB	5 dB	0 dB	-5dB
MFCC-SM2-NORM	91.54	74.50	61.51	45.09	28.19	16.95	11.11
LOG-DCT-SM2-NORM	95.72	84.10	76.54	64.76	47.42	27.03	12.61
HC-DCT-SM2-NORM	94.29	92.00	88.96	80.91	63.28	35.28	10.56
PAS-SM2-NORM	95.19	93.67	90.62	81.00	61.64	27.55	1.73

Abbildung 9.3 ***Erkennraten der VAMIG Ausprägungen ohne Schwellenanpassung, mit Schwellenanpassung und Normalisierung bei einer unteren Grenzfrequenz von 200 Hz***

	INF	20 dB	15 dB	10 dB	5 dB	0 dB	-5dB
MFCC-SM0-75Hz	97.36	80.54	65.51	45.01	22.41	8.58	5.57
LOG-DCT-SM0-75Hz	95.96	72.09	59.04	42.18	20.33	9.69	7.78
HC-DCT-SM2-75-Hz	93.94	88.76	77.02	49.07	20.40	8.57	5.50
PAS-SM2-75Hz	95.85	92.74	85.75	68.77	39.20	14.93	7.87

Abbildung 9.4 ***Erkennraten der jeweils besten VAMIG- Ausprägungen bei einer Grenzfrequenz von 75 Hz, ohne Schwellenanpassung und ohne Normalisierung.***

	INF	20 dB	15 dB	10 dB	5 dB	0 dB	-5dB
MFCC-SM0	97.36	80.54	65.51	45.01	22.41	8.58	5.57
PAS-SM2	95.85	92.74	85.75	68.77	39.20	14.93	7.87
PAS-CAS-PCF-SM2	94.73	93.19	89.87	77.67	50.64	20.86	8.98
PAS-CAS-KM-SM2	95.04	93.47	88.38	72.28	42.57	17.81	9.11

Abbildung 9.5 ***Vergleich der Erkennraten der jeweils besten VAMIG- Ausprägungen bei einer unteren Grenzfrequenz von 75 Hz, die Erkennraten für die Normalisierung wurden hier nicht berücksichtigt.***

	INF	20 dB	15 dB	10 dB	5 dB	0 dB	-5dB
MFCC-SM2-NORM							----
PAS-SM2-NORM	95.67	93.67	90.62	81.00	61.64	27.55	----
PAS-CAS-PCF-SM2-NORM	94.02	93.17	90.82	83.70	64.41	30.82	----
PAS-CAS-KM-SM2-NORM	94.71	93.40	89.76	78.99	56.01	26.72	----

Abbildung 9.6 ***Vergleich der Erkennraten der jeweils besten VAMIG- Ausprägungen bei einer unteren Grenzfrequenz von 75 Hz und eingeschalteter Normalisierung.***

In einem weiteren Experiment wurden schließlich die Erkennraten für die hochauflösenden *PAS-CAS*-Modelle aus Abschnitt 5.4 ermittelt. In diesem Zusammenhang soll noch einmal daran erinnert werden, dass bei diesen Modellen sowohl die Geräuschreduktion als auch die Bindung im Frequenzbereich (Absolutbetrag des Spektrums) realisiert wurde. Bei allen anderen Modellen erfolgte die Geräuschreduktion (Schwellenanpassung) und auch die Bindung auf der *RMS*-Ebene.

	INF	20 dB	15 dB	10 dB	5 dB	0 dB	-5dB
MFCC-SM0	97.36	80.54	65.51	45.01	22.41	8.58	5.57
PAS-SM2	95.85	92.74	85.75	68.77	39.20	14.93	7.87
PAS-CAS-PCF-SM2	94.73	93.19	89.87	77.67	50.64	20.86	8.98
PAS-CAS-KM-SM2	95.04	93.47	88.38	72.28	42.57	17.81	9.11
HA-PAS-CAS-PCF-SM2	94.16	92.69	89.20	80.16	59.50	30.65	10.20
HA-PAS-CAS-PRED-SM2	94.87	93.31	90.00	80.11	59.69	30.76	10.52

Abbildung 9.7 ***Vergleich der mittleren Erkennraten der besten VAMIG- Ausprägungen bei Hinzunahme der hochauflösenden PAS-CAS-Modelle bei einer unteren Grenzfrequenz von 75 Hz.***

Die folgende Tabelle zeigt die letzten Ergebnisse noch einmal in einem anderen Kontext. In der ersten Zeile findet man die Ergebnisse mit dem Modell *MFCC-Aurora-SM0*, welche als Referenz im Aurora-2 Set zur Verfügung gestellt wurde. Um Vergleichbarkeit zu gewährleisten, wurden die *Δ* und *ΔΔ* Koeffizienten auch hier nur mit der ersten und zweiten Differenz jeweils erster Ordnung bestimmt. In der zweiten Zeile wurde für das einfache *PAS*-Modell (ohne Bindung) die Spektrale Subtraktion ebenfalls in den Frequenzbereich verlegt, auf die Schwellennormalisierung in der *RMS*-Ebene konnte daher verzichtet werden. Die dritte Zeile zeigt die Ergebnisse für das gleiche Modell nun aber bei Normalisierung der Supervektor-Komponenten. Die letzten beiden Zeilen zeigen noch einmal die Erkennraten für die hochauflösenden *PAS-CAS*-Modelle.

	INF	20 dB	15 dB	10 dB	5 dB	0 dB	-5dB
MFCC-AURORA-SM0	98.86	92.80	81.43	61.93	35.10	13.80	7.70
HA-PAS-SM2	95.85	89.52	75.79	51.09	24.09	11.16	10.29
HA-PAS-NORM-SM2	96.13	93.97	87.56	72.78	46.27	13.88	6.71
HA-PAS-CAS-PCF-SM2	94.16	92.69	89.20	80.16	59.50	30.65	10.20
HA-PAS-CAS-PRED-SM2	94.87	93.31	90.00	80.11	59.69	30.76	10.52

Abbildung 9.8 ***Vergleich der mittleren Erkennraten der hochauflösenden PAS-CAS-Modelle (mit Bindung) mit der Aurora-2 Vorverarbeitung und mit den hochauflösenden PAS-Modellen (ohne Bindung).***

In der letzten Tabelle werden die Erkennraten aller relevanten Modelle nochmals über alle SNR gemittelt. Bis auf die zweite Zeile ist die untere Grenzfrequenz immer 75 Hz, von der Aurora-2 Vorverarbeitung ist dieser Parameter leider nicht bekannt.

	MEAN
MFCC-SM0	46.43
MFFC-AURORA-SM0	55.94
PAS-SM2	57.87
PAS-CAS-PCF-SM2	62.28
PAS-CAS-KM-SM2	59.81
HA-PAS-CAS-PCF-SM2	65.22
HA-PAS-CAS-PRED-SM2	65.61

Abbildung 9.9 ***Vergleich der mittleren Erkennraten der besten VAMIG- Ausprägungen, nun bei zusätzlicher Mittelung über alle SNR.***

Bewertung der Erkennergebnisse

Zunächst kann beobachtet werden, dass lediglich die auditiven Modelle mit dynamischer Kompression von der Schwellennormalisierung profitieren. Die Anwendung der Schwellennormalisierung bei statischer Kompression führt zu einem drastischen Einbruch der Erkennraten. Dagegen führt eine Normalisierung bei allen *VAMIG*-Ausprägungen zu deutlichen Verbesserungen. Die *PAS-CAS-Modelle* zeigen ab 20 dB Störabstand ebenfalls deutliche Verbesserungen sowohl gegenüber dem *MFCC*-Standardmodell als auch gegenüber dem *PAS-Modell*.

Die Integration der Kalmanvorhersage führt nur bis etwa 20 dB zu leichten Verbesserungen, danach fällt diese Ausprägung gegenüber der strukturell verbesserten *PCF*-Ausprägung wieder ab. Die gemessenen 97.36 % bzw. 98.32 % der *MFCC*-Modelle erreicht allerdings keines der auditiven Modelle.

Über den gesamten SNR-Bereich gesehen zeigen die rein auditiven Ausprägungen die besseren Erkennraten. Allerdings führen bereits relativ einfache Maßnahmen wie Preemphasis, optimales Flooring und eine höhere Ordnung bei der Berechnung der Δ und ΔΔ Koeffizienten auch im *MFCC*-Modell zu Verbesserungen oberhalb von 20 dB. Andererseits kann durch eine modifizierte Parametrierung (siehe Änderung der unteren Grenzfrequenz) der *PAS-CAS*-Modelle weiteres Optimierungspotential erschlossen werden. Insbesondere eine erhöhte Abtastfrequenz im Verbund mit einer Erhöhung der Anzahl der Teilbänder könnte eine robustere Schätzung der Sprachgrundfrequenz und damit der Gewichtungsvektoren zur Folge haben. Für den Bereich der guten bis sehr guten SNR müssen allerdings weitere Anstrengungen unternommen werden, hier ist die absolute Differenz zwischen den *PAS-CAS*-Modellen und den *MFCC*-Modellen von etwa 2% bzw. 4% unbefriedigend.

Die Implementierung von Geräuschreduktion und Bindung im Frequenzbereich hat zu einer weiteren Verbesserung der Erkennraten geführt. Durch die Integration der Vorhersage konnten, im Gegensatz zur Implementierung in der *RMS*-Ebene, nun auch die erhofften Auswirkungen auf die Erkennrate beobachtet werden. Da diese Realisierungsform darüber hinaus auch die recheneffizientere Variante ist (es ist nur noch eine *IFFT* zur Berechnung *AKF* notwendig), scheint das hochauflösende *PAS-CAS*-Modell selbst für Embedded-Systeme mit geringen Ressourcen eine geeignete Wahl zu sein *[Römer-00]*.

Um sicherzustellen, dass die Ursache für die Verbesserungen bei den hochauflösenden Modellen tatsächlich in der Bindung und nicht etwa nur in einer wirksameren Spektralen Subtraktion (in der Frequenzebene, nicht in der RMS-Ebene) zu finden ist, wurden die Erkennraten in Abbildung 9.8 gemessen. Dabei wurde die Spektrale Subtraktion nun auch für das *PAS*-Modell in den Frequenzbereich verschoben. Zusätzlich wurden zum Vergleich auch noch die Erkennraten der Aurora-2-Vorverarbeitung herangezogen. Vergleicht man nun die Erkennraten miteinander, so kommt man zu dem Schluss, dass sowohl die Bindung als auch die Einbindung der Vorhersage zu einer weiteren Verbesserung der Erkennraten führen.

In Abbildung 9.9 wurden die Erkennraten nochmals zusätzlich über alle SNR gemittelt, dabei treten die Unterschiede noch deutlicher hervor. Der relative Unterschied zwischen dem besten *MFCC*-Modell *MFCC-AURORA-SM0* und dem besten hochauflösenden *PAS-CAS*-Modell beträgt hier 21.94 %. Selbst wenn man das unveränderte Setting der Aurora-2 Vorverarbeitung aus *[Hirsch-00]* übernimmt (gemittelte Erkennrate über alle SNR: 59.09 %), was eigentlich nicht direkt vergleichbar ist, kann man gegenüber dem besten hochauflösenden *PAS-CAS*-Modell immer noch eine relative Verbesserung von etwa 16 % beobachten. Hier beginnt der Gewinn ab einem SNR von 15 dB.

Ebenfalls von Interesse ist ein Vergleich zwischen Modellen welche nur *PAS*-Eigenschaften realisieren und Modellen bei denen zusätzlich auch *CAS*-Eigenschaften realisiert worden sind. Dabei kann festgestellt werden, dass die *PAS-CAS*-Modelle bei additiven Störungen die besten Ergebnisse erzielen und sich daher „robuster“ verhalten als die einfachen *PAS*-Modelle. Mit Tabelle 9.10 kann diese Feststellung auch quantitativ noch einmal untermauert werden.

	MEAN	Rel. Gewinn
PAS-SM2	57.87 %	…..
PAS-CAS-PCF-SM2	62.28 %	10.47 %
PAS-CAS-KM-SM2	59.81 %	4.59 %
HA-PAS-CAS-PCF-SM2	65.22 %	17.42 %
HA-PAS-CAS-PRED-SM2	65.22 %	18.34 %

Abbildung 9.10 ***Mittlerer relativer Gewinn der PAS-CAS-Modelle gegenüber dem besten PAS-Modell. Hier schneiden die hochauflösenden Modelle, welche stärker an das Redundanz- und Kontinitäts-prinzip gebunden sind, am besten ab.***

Gezeigt wird hier der mittlere relative Gewinn der einzelnen *PAS-CAS*-Modelle gegenüber dem besten *PAS*-Modell, gemessen über allen Störgeräuschen und Störintensitäten. Das dieser Gewinn nicht auf unterschiedliche Methoden der Geräuschunterdrückung zurückzu-führen ist, wurde bereits in Tabelle 9.8 gezeigt.

10 Schlussfolgerungen und Ausblick

Die vorliegende Arbeit wurde stark von den Ausführungen in *[Allen-94]* beeinflusst. Die wesentlichen Kernaussagen (lokale Informationsverarbeitung in den Teilbändern, Bindung von sich gleichsinnig entwickelnden Signalkomponenten und Mechanismen zur Verbesserung des lokalen SNR) sollen nun vor dem Hintergrund der vorliegenden Ergebnisse beleuchtet werden. Zunächst muss man feststellen, das eine echte lokale Signalverarbeitung von Teilbändern nur für das *PAS* realisiert worden konnte (Wavelet-Zerlegung). Für die Integration der Bindungseigenschaft fehlte allerdings noch ein geeigneter Algorithmus.

Alle anderen Modelle verwenden zunächst eine *STFT,* erst die anschließende Filterbank erzeugt Teilbänder, in denen dann eine lokale Informationsverarbeitung erfolgt. Dies gilt auch für die Vorhersagen. Die Bindung erfolgt (lokal in der Zeitdimension) in den stimmhaften Abschnitten, dabei wird das SNR (lokal in der Frequenzdimension) nur für diejenigen Teilbänder verbessert, bei denen die Einhüllende von der Sprachgrundfrequenz dominiert wird. Darüber hinaus konnten lokale SNR-Verbesserungen für das Haarzellen-Modell festgestellt werden. Diese Verbesserungen könnten noch sehr viel deutlicher ausfallen, dazu müssten die Haarzellenmodelle allerdings in mehreren Stufen und mit unterschiedlichen Zeitkonstanten ausgeführt werden *[Tchorz-99]*. Für solche Modelle wird dann allerdings eine echte Teilbandverarbeitung notwendig sein, bei der man wohl auch mit der vollen Abtastrate arbeiten müsste. Eine Alternative könnte hier wieder mit der Wavelet-Zerlegung vorliegen, dann müssten die Haarzellenmodelle allerdings mit unterschiedlichen Abtastfrequenzen betrieben werden.

Aus praktischen Gründen war es für die Erkennexperimente (*HMM*-Technologie) notwendig, eine abschließende *DCT* zur Dekorrelation der Merkmale zu berechnen. Da sowohl die *STFT* als auch die *DCT* mit globalen Basisfunktionen arbeitet, können die Merkmale nicht unabhängig voneinander sein (impulsartige Signalabschnitte werden auf alle Frequenzen verteilt). D.h. wenn man auditive Modelle verwendet, in denen eine echte Teilbandverarbeitung stattfindet, dann wird man wahrscheinlich auch die *HMM*-Klassifikatoren an solche Modelle anpassen müssen. Andererseits könnte man natürlich auch Neuronale Netze oder andere Ansätze (siehe Ausblick) verwenden.

Eine wichtige Erweiterung auditiver Modelle stellt die Einbindung von Modulationsfilterbänken dar, hierbei würde jedes Teilband nochmals einer Filterbank zugeführt werden, und hinsichtlich der auftretenden Frequenz- und Amplitudenmodulationen analysiert werden. Erste Modelle sind hier bspw. von *[Chi-03],[Mesgerani-05]* und *[Dau-99]* präsentiert worden. In der Automatischen Spracherkennung scheitert der Einsatz solcher Modelle derzeit noch an den immensen Rechenzeit- und Speicheranforderungen.

10.1 Bewertung der Hypothese zur Integration der Bindungseigenschaft

In Kapitel 9 wurde dokumentiert, dass bei mittleren und schlechten *SNR* mit den rein auditiven *VVM* und auch den hochauflösenden *PAS-CAS*-Modellen höhere Erkennraten erzielt werden, als das für die konventionellen *VVM* der Fall ist. Für diesen *SNR*-Bereich wurde erwartet, dass die auditiven Modelle mit integrierter Bindungseigenschaft die Erkennraten weiter steigern können. Da die beobachtete Verbesserung sehr deutlich hervor getreten ist, kann die Hypothese zur Integration der Bindungseigenschaft recht sicher bestätigt werden. Man kann also durchaus davon ausgehen, dass die Bindungseigenschaft einen wichtigen Beitrag zur Zuverlässigkeit auditiver Merkmale liefert und daher in einer auditiv motivierten Merkmalsextraktion eingesetzt werden sollte. Darüber hinaus sollte erwähnt werden, dass die Integration der Bindungseigenschaft überhaupt nicht auf die auditiven Modelle beschränkt ist.

Insbesondere der Ansatz der hochauflösenden *PAS-CAS*-Modelle kann durchaus auch auf das *MFCC*-Modell übertragen werden.

Auf einen weiteren Aspekt soll in diesem Zusammenhang ebenfalls noch hingewiesen werden: Von der Schwellennormalisierung profitieren nur die auditiven Modelle mit dynamischer Kompression, dagegen profitieren alle VAMIG- Ausprägungen sehr stark von der Komponenten-Normalisierung. Dabei ist nun allerdings zu berücksichtigen, dass die Information zur Normalisierung erst dann zugänglich wurde, wenn die Äußerung bereits vollständig verfügbar war. Dies ist in Echtzeit kaum zu realisieren. Interessant erscheint nun, dass die Komponenten-Normalisierung bei den *PAS-CAS*-Modellen nur noch zu leichten Verbesserungen führte, obwohl die Information zur Bindung lediglich vom aktuellen Frame gewonnen werden kann (siehe Tabelle 9.5 und Tabelle 9.6). Insbesondere bei schlechten SNR (etwa ab 10 dB) lassen sich aber bei einer Kombination der beiden Verfahren nochmals deutliche Gewinne erzielen. Man muss dann allerdings Abstriche bei der Echtzeitfähigkeit in Kauf nehmen.

In *[Viiki-97]* und *[Viiki-98]* wurde berichtet, dass die Komponenten-Normalisierung auch mit relativ kleinen Normalisierungsfenstern adaptiv durchgeführt werden kann. Dieses Verfahren könnte man demnach mit der bedingten Bindung kombinieren, ohne die Echtzeitfähigkeit in Frage stellen zu müssen.

10.2 Bewertung der Hypothese zur Integration von Vorhersagen

Wenn man die Ergebnisse aus Tabelle 9.5 zur Bewertung der zweiten Hypothese heranzieht, so kann man für die Realisierung in der *RMS*-Ebene zunächst keinen signifikanten Fortschritt feststellen. Bei sehr gutem *SNR* kann man zwar eine leichte Erhöhung der Erkennrate beobachten, allerdings ist dann für mittlere und schlechte *SNR* wieder ein Abfall gegenüber dem einfachen Bindungsmodell zu beobachten. Ein anderes Bild zeigte sich dagegen für das hochauflösende *PAS-CAS*-Modell: Über alle *SNR* gemittelt wurden mit diesem Modell die besten Ergebnisse erzielt, auch wenn der Zuwachs nicht mehr besonders stark ausfiel. Aus diesem Grund erscheint die Integration von Vorhersagen auf der Merkmalsebene vielleicht nicht unbedingt nötig. Möglicherweise kann die Integration von Vorhersagen ihre volle Wirksamkeit aber erst auf höheren Sprachverarbeitungsebenen entfalten. Auf diesen Aspekt, der das im Vorwort erwähnte Ökonomieprinzip nun in den Mittelpunkt rückt, wird im folgenden Ausblick genauer eingegangen.

10.3 Ausblick

Die Ideen zur Integration der Bindungseigenschaft sind nicht neu, sie wurden in *[Fletcher-53]* bereits vorweggenommen und nun auf der Merkmalsebene umgesetzt. Problematisch erscheint die Leistungsfähigkeit auditiver Modelle bei sehr guten SNR. Insbesondere bei den *PAS-CAS*-Modellen muss man Wege finden, um die für die Linearkombination wichtige Stimmhaftigkeit besser schätzen zu können. Neben dem Verbesserungspotential auditiver Modelle bei sehr guten *SNR* sollten auch die höheren Sprachverarbeitungsebenen von den neueren Erkenntnissen über Struktur und Funktionalität des menschlichen Gehirns profitieren.

In den letzten Jahren sind nun verstärkt Bemühungen (hauptsächlich in der Bildverarbeitung) unternommen worden, um weitere strukturelle Eigenschaften des Großhirnrinde (Neocortex) in Modellen zur Merkmalsextraktion zu verankern *[Mumford-03].* Dabei kommt denjenigen Modellen eine besondere Bedeutung zu, bei denen die Merkmalsextraktion auf mehrere hierarchisch geordnete Schichten verteilt ist. Bei diesem Ansatz kommt das Vorhersageprinzip konsequent in allen Ebenen des Schichtenmodells zur Anwendung.

Von besonderer Bedeutung ist dabei der Übergang von der subsymbolischen Signalverarbeitung (Schwerpunkt dieser Arbeit) zur symbolischen Informationsverarbeitung (hier kann der Kontext eingebunden werden). Dieser Schritt ist eine notwendige Voraussetzung für die Anwendung von hierarchischen Modellen. Im Folgenden werden beispielhaft zwei Modelle vorgestellt, mit denen die Vorteile der hierarchischen Informationsverarbeitung deutlich gemacht werden können und die darüber hinaus den Weg für zukünftige Arbeiten weisen.

10.3.1 Das Gedächtnis-Vorhersage-Modell nach J. Hawkins

In *[Hawkins-04]* wird auf eine Theorie zur Funktionsweise der Großhirnrinde (Neokortex) von V. Mountcastle hingewiesen, die lange Zeit ignoriert wurde. Zum besseren Verständnis dieser Theorie wird zunächst auf den Aufbau des Neokortex eingegangen:

- Die Großhirnrinde besteht aus 6 Ebenen.
- Die einzelnen Regionen der sechs übereinander liegenden Ebenen lassen sich in so genannte “Säulen” aus übereinander liegenden Zellen einteilen, die durch Axone miteinander verbunden sind. Diese Säulen können als Basis-Einheiten des Neokortex verstanden werden.
- Die Einteilung der Großhirnrinde in verschiedene Regionen ist hierarchisch organisiert.
- Der Informationsfluss ist bidirektional, d.h. sowohl von niedrigeren Regionen zu den höheren als auch umgekehrt (Feedback).

Die nach Hawkins wichtigste Entdeckung der Neurowissenschaften ist Mountcastles These, dass die verschiedenen Regionen der Großhirnrinde grundsätzlich nach dem gleichen Prinzip – einem Ökonomieprinzip – funktionieren *[Mountcastle-78]*. Diese Regionen führen sozusagen alle den gleichen Algorithmus aus. Dazu kommt, dass sich die Eingaben, die die Großhirnrinde erreichen, sehr stark ähneln: Hören, Sehen und andere Sinne werden alle über Nervenleitungen ans Gehirn geschickt (Spikefolgen). Es handelt sich also um nichts anderes als zeitliche und räumliche Muster. Die Erkennung von solchen Mustern geschieht nach Hawkins nun nicht über aufwendige neuronale Berechnungen, sondern durch den massiven Einsatz eines autoassoziativen Speichersystems. Hawkins gliedert seine These wie folgt:

- Die Großhirnrinde speichert Sequenzen von Mustern,
- Die Großhirnrinde ruft die Muster autoassoziativ ab, d.h. teilweise vorhandene räumliche / zeitliche Muster aktivieren das komplette Muster.
- Die Großhirnrinde speichert Muster in einer invarianten Form, d.h. nur die wichtigen Zusammenhänge der Welt werden gespeichert, nicht die Details einer Momentaufnahme.
- Die Großhirnrinde speichert Muster hierarchisch.

Durch die Kombination aus invarianten Repräsentationen und den Informationen der Sinnesorgane über die momentane Situation kann das Gehirn Vorhersagen über aktuelle Geschehnisse machen. Diese Vorhersagen spielen eine entscheidende Rolle um etwas zu “verstehen”.

Demnach setzt sich die Wahrnehmung der Welt nicht nur aus den mit unseren Sinnen empfangenen Informationen, sondern zugleich aus den ununterbrochen erstellten Vorhersagen über die Struktur der Welt zusammen. Das Gehirn erstellt also ein Modell der Realität und vergleicht dieses ständig mit den tatsächlichen Sinnesempfindungen.

Diese Vorhersagen finden sowohl in den niedrigsten Regionen, die die Eingaben der Sinneszellen erhalten, als auch in den höheren Regionen des Denkens statt. Sie sind die primäre Funktion des Algorithmus der Großhirnrinde. Während sich in den niedrigeren Regionen der Großhirnrinde die Neuronenaktivität ständig ändert, findet sich umso mehr konstante Aktivität, je "höher" der Status einer Region. Es gibt beispielsweise Neuronen, die immer dann feuern, wenn wir ein Gesicht sehen. Diese Tatsache bezeichnet Hawkins als invariante Repräsentationen. Während niedrigere Regionen sinnesspezifisch sind, integrieren höhere Regionen die verschiedenen Sinne: Ein Geräusch kann auf einer höheren Ebene die Repräsentation eines Objekts aktivieren, mit deren Hilfe auf einer niedrigeren visuellen Ebene Vorhersagen gemacht werden können, was wir als nächstes Sehen werden.

Die Aufgabe jeder Region ist es herauszufinden, in welcher Beziehung (Syntax) die einzelnen Signale einer Region zueinander stehen. Wird dann noch die Reihenfolge ihres Auftretens gespeichert, so kann man vorherzusagen, welche Signale zukünftig zu erwarten sind. Jede Region erzeugt also invariante Repräsentationen anhand der aus den niedrigeren Regionen kommenden Daten.

Die Großhirnrinde ist hierarchisch aufgebaut, d.h. von den sinnesspezifischeren niedrigeren Regionen hin zu den abstrakt denkenden, höheren Regionen. Zu jedem Zeitpunkt können wir aber nur einen kleinen Teil einer Situation mit den Sinnen erfassen: Durch den hierarchischen Aufbau der Großhirnrinde wissen wir dennoch, wie dieser Moment Teil einer übergeordneten Situation ist, da invariante Repräsentationen in höheren Regionen die ganze Zeit über aktiv bleiben. Demnach ist unser Verständnis der Realität vom Prinzip der Vorhersagbarkeit geprägt: Wenn das Gehirn das nächste Muster einer Sequenz akkurat vorhersagen kann, wird das bei mehrmaligem Auftreten als kausale Beziehung interpretiert. "Erkennt" eine Gehirnregion eine Sequenz, so wird der "Name" (nicht das eigentliche Muster) dieser Sequenz an die nächst höhere Region weitergegeben so lange diese Sequenz aktiv ist. Das Gehirn kann so Sequenzen von Sequenzen erkennen und speichern. Je höher eine Gehirnregion, desto mehr Konstanz ist zu beobachten, da übergeordnete Sequenzen länger aktiv bleiben.

Da die Anzahl an möglichen Kombinationen der Eingabesignale für eine Gehirnregion unermesslich groß ist, stellt sich die Frage wie diese Sequenzen entstehen. Die Antwort lautet: Jede Gehirnregion ordnet die Signale die es empfängt, einem bestimmten Cluster zu. Versieht man die Cluster mit einer Bedeutung, so entstehen Symbole die dann in einem Alphabet zusammengefasst werden können. Sequenzen bestehen demnach aus symbolischen Zuordnungen. Lässt sich eine Eingabe nicht eindeutig einem Symbol zuordnen, so wird der Kontext der aktuellen Sequenz als Hilfe zugezogen und die Eingabe mit dem erwarteten Symbol verglichen. Höhere Regionen der Großhirnrinde geben hierzu Informationen an niedrigere Regionen weiter, welche Eingabe sie als nächstes zu „erwarten" haben.

Einzelne Regionen der aus sechs übereinander liegenden Ebenen bestehenden Großhirnrinde lassen sich in so genannte "Säulen" aus übereinander liegenden Zellen einteilen, die durch Axone verbunden sind und meist zur gleichen Zeit aktiv werden. Diese Säulen können als Basis-Einheiten des Vorhersage-Modells betrachtet werden. Die große Anzahl an Synapsen, die die Zellen einer Säule mit Zellen aus anderen Teilen des Gehirns verbindet, sorgt für die nötigen Kontext-Informationen. Informationen in den Säulen fließen dabei von unten nach oben direkt (konvergierend) während sie von oben nach unten viele Verzweigungen nehmen müssen. Auf diese Weise werden invariante Repräsentationen aus höheren Regionen in spezifische, der Situation angepasste Signale verwandelt.

In jüngerer Zeit sind nun einige Modelle entwickelt worden, mit denen die oben beschriebene Theorie in Simulationen evaluiert werden kann *[George-05],[George-06].* Den Kern dieser Modelle bildet der so genannte ***B**elief-**P**ropagation-**A**lgorithmus* (*BPA*).

Belief- Propagation

Der Belief-Propagation-Algorithmus beruht auf den sogenannten zerlegbaren graphbasierten Modellen, die auch als Bayes'sche Netze oder Belief-Netze bekannt sind *[Pearl-88]*. Eine charakteristische Eigenschaft dieser Netze ist, dass Unsicherheit und Vertrauen in einem Netzwerk auf der Basis ausschließlich lokaler Operationen propagiert werden: Jeder Knoten des Netzes wird als eigenständiger Prozessor angesehen, der mit seinem Nachbarknoten Nachrichten austauscht.

Ein bei der Verwendung von Netzwerken zur Handhabung von Unsicherheit grundlegendes Konzept ist die Repräsentierbarkeit von Abhängigkeiten und Unabhängigkeiten. Mit diesem Konzept gelingt es, die Anzahl möglicher Symbolkombinationen deutlich zu reduzieren. Das ist eine Grundvoraussetzung für die Verwendung von hierarchischen Strukturen.

Die Funktion des *BPA* für ein baumartig strukturiertes Knoten-Netzwerk kann man sich folgendermaßen vorstellen: Die Knoten in der Sensorebene triggern die Knoten in den darüber liegenden Schichten sobald die Sensoren Änderungen in den Eingangssignalen feststellen. Dabei werden Nachrichten in Form von WK-Verteilungen immer nur lokal nach oben propagiert (Diagnostik). Wenn ein Top Knoten von darunter liegenden Knoten Nachrichten erhält, kann er eine Entscheidung über das Toplevel Symbol treffen. Ausgehend von diesem Symbol kann nun die WK-Verteilung für das nächste Toplevel Symbol vorhergesagt werden und nach unten propagiert werden (Prädiktion). Da nun jeder Knoten sowohl über diagnostische als auch über prädiktive Nachrichten verfügt, kann in jedem Knoten eine so genannte Beliefverteilung berechnet werden. Diese Verteilung ergibt sich durch die Berücksichtigung aller zur Verfügung stehender Wissensquellen (diagnostische und prädiktive WK-Verteilungen).

10.3.2 Die Sprachmaschine nach W. Hilberg

Ein ähnlicher Ansatz wird von W. Hilberg bei der hierarchischen Textkomprimierung verfolgt. Hilberg orientiert sich am Zipfschen Gesetz, wonach die Häufigkeit der Worte aufgetragen über der Rangfolge von Worten für alle natürlichen Sprachen eine Gerade mit einer Steigung von etwa „-1" ist. Hilberg vermutet, dass diese Gerade Ausdruck einer allen Menschen gemeinsamen Gedächtnisstruktur – also einer Invarianz – ist *[Hilberg-97a], [Hilberg-97b]*. Darüber hinaus zeigen die menschlichen als auch die von Hilberg künstlich erzeugten Sprachnetzwerke ein für die Evolution typisches Verhalten, nämlich das der maximalen Entropie bzw. der maximalen Informationsspeicherung. Typisch für diesen Ansatz ist dabei die Verwendung eines hierarchischen Gedächtnissystems. Innerhalb eines solchen Systems erfolgt die Informationsverarbeitung sowohl auf vertikaler als auch auf horizontaler Ebene. D.h. neben der Beobachtung von zulässigen Symbolfolgen in der Basisebene, werden auch zulässige Symbolfolgen in den Metaebenen erfasst. Mit dieser zusätzlichen Information können tiefer liegende Metaebenen gesteuert werden. D.h. neben den syntaktischen Aspekten beeinflussen auch strukturelle Aspekte die Leistungsfähigkeit hierarchischer Systeme.

Strukturelle Aspekte

Bei W. Hilberg wird ebenfalls ein hierarchisches Schichtensystem als Gedächtnisstruktur vorgeschlagen, in dem die Zulässigkeit des Aufeinanderfolgens von immer größer werdenden Metawörtern gespeichert ist.

Das hierarchische Gedächtnissystem besteht aus strukturell gleichartigen Netzwerken, die nach oben hin einen immer stärkeren Abstraktionsgrad aufweisen. In einem solchen Gedächtnissystem werden jeweils zwei aufeinander folgende Worte zu einem Metawort zusammengefasst und unter einem neuen Namen in der nächst höheren Schicht gespeichert. Die zulässigen Wortübergänge werden in sogenannten Assoziationsmatrizen gespeichert. Diese Prozedur wiederholt sich für alle Metaebenen, so dass in der obersten Metaebene letztlich nur noch ein einziges Metawort übrig bleibt. Hilberg hat gezeigt, dass Wortfolgen auf diese Weise fast völlig redundanzfrei codiert werden können. Der umgekehrte Prozess, also das Entfalten des obersten Metaworts in eine Wortfolge wird Decodieren genannt, dabei steuert das Metawort den Textpfad der nächst tieferen Ebene. Jede Metaebene verfügt dann aber über ein eigenes (immer größer werdendes) Alphabet. Ohne geeignete Maßnahmen explodiert ein solches System hinsichtlich der möglichen Symbolkombinationen.

Innerhalb dieser Gedächtnisstruktur findet also sowohl eine horizontale (Durchschreiten gerichteter Verbindungen innerhalb der Netzwerkpfade) als auch eine vertikale Informationsverarbeitung (Steuerung der Netzwerkpfade) statt. Die horizontale Informationsverarbeitung beschreibt im Wesentlichen zeitliche Abläufe innerhalb eines Schichtennetzwerks. Die vertikale Informationsverarbeitung berücksichtigt dagegen die strukturellen Aspekte. Prädiktionen innerhalb eines Schichtennetzwerks stellen demnach nur einen von zwei Gedächtnisaspekten (Implizite Prädiktion) dar. Der zweite Gedächtnisaspekt lässt sich aus der vertikalen Sprachverarbeitung ableiten (Explizite Prädiktion).

10.3.3 Codieren und Decodieren

Im Allgemeinen „explodieren“ in einem hierarchischen System die Alphabete der höheren Schichten. Mit der Sprachmaschine nach W. Hilberg wird gezeigt, dass man die Anzahl der Symbole der verschiedenen Schichtenalphabete konstant halten kann. Die Reduktion dieser Alphabete auf das Alphabet der Basisebene gelingt technisch nur unter Verwendung der Prinzipien: Ähnlichkeitsbündelung (Cluster), Verdichtung und implizite Prädiktion. Mit diesen Prinzipien lassen sich Texte codieren und decodieren, dabei wird die Syntax (Grammatik) implizit im hierarchischen Netzwerk gespeichert. Bei der Codierung werden immer ein Leitwort und das zugehörige Ähnlichkeitsbündel (z.B. aus der Assoziationsmatrix) zu einem Metawort der nächst höheren Ebene zusammengefasst, der Name des Metaworts ist identisch mit dem Namen des Leitworts. Demnach bleibt das Alphabet aller höheren Schichten identisch zum Basisalphabet.

Das oberste verbleibende Metawort stellt dann eine redundanzfreie Darstellung des Textes – also einen Gedanken oder eine Idee – dar. Als Leitworte werden die Worte mit dem höheren Informationsgehalt, also die selteneren Worte gewählt. Diese Leitworte können bei der Decodierung direkt als Stützstellen für die implizite Prädiktion genutzt werden. Die Entfaltung einer Idee führt wieder auf den ursprünglichen Text. Da neben dem Leitwort zunächst nur ein Ähnlichkeitsbündel zur Decodierung vorliegt, wird das Hilfsmittel der Prädiktion verwendet um Mehrdeutigkeiten aufzulösen. D.h. die Prädiktionen auf den Ähnlichkeitsbündeln kann nur innerhalb einer hierarchischen Ebene erfolgen und zwar auf der Basis bekannter Vorgänger und Nachfolger.

10.3.4 Vergleich der beiden Modelle

Ein Vergleich der Modelle von J. Hawkins und W. Hilberg zeigt, dass sich beide Verfahren vor allem in den Verwendungsmöglichkeiten unterscheiden: Hawkins konzentriert sich auf Mustererkennungen, wobei die beobachteten Muster prinzipiell gestört sein können. Zur Decodierung sind somit statistische Verfahren notwendig.

Hilberg dagegen schlägt ein System zur Textkomprimierung vor, bei dem die Texte ungestört sind. Gleiches gilt für die Gedächtnisstruktur, diese ist ebenfalls determiniert. Beide Ansätze unterscheiden sich aber auch hinsichtlich der Verwendung des Vorhersageprinzips:

Bei Hawkins werden zwischen den Schichten sogenannte ***C**onditional **P**robability **T**ables* (*CPT*) verwendet, mit diesen Matrizen erfolgte eine Beschreibung der Beziehung zwischen den Symbolen von Alphabeten von zwei aufeinander folgenden Ebenen. Diese Matrix kann als Prädiktor in vertikaler Ebene verstanden werden. Dagegen beschreibt die Assoziationsmatrix bei Hilbert die Beziehungen zwischen aufeinanderfolgenden Symbolen des Alphabets innerhalb einer Ebene. Diese Matrix kann als Prädiktor in horizontaler Ebene verstanden werden.

10.3.5 Verallgemeinerte hierarchische Informationsverarbeitung

Die Merkmalsextraktion wird häufig als selbständige Verarbeitungsstufe zwischen Vorverarbeitung und Klassifikation eingeschoben. Das Ziel besteht dann darin, Merkmale mit möglichst geringer Dimensionalität zu finden, welche die relevanten Eigenschaften der Muster enthalten und somit den Entwurf des Klassifikators erleichtern. In den meisten Fällen wird man sich allerdings mit dem Problem konfrontiert sehen, dass die signifikanten Merkmale nicht bekannt sind. Für die Merkmalsvektoren wird man dann in einem ersten Schritt möglichst viele Komponenten vorsehen. Mit einer nachfolgenden Merkmalsselektion können anschließend die wichtigsten Komponenten ausgewählt werden.

In der Sprachverarbeitung kommt häufig die Lineare Diskriminanzanalyse (*LDA*) zum Einsatz *[Ruske-94]*. Dabei erfolgt nun nicht nur eine einfache Auswahl der wichtigsten Komponenten, sondern mit der LDA werden auch die Forderungen erfüllt, dass die Komponenten voneinander dekorreliert sind und bei gegebener Dimensionsreduktion die maximale Trennbarkeit zwischen den Klassen sicher gestellt wird. Eine solche Merkmalsextraktion nennt man auch Transformations-Codierung. Dagegen wird in biologischen Systemen eher das Prinzip der spärlichen Codierung bevorzugt *[Burke-97]*. Hier werden nicht einige wenige Komponenten für einen Merkmalsvektor ausgewählt, welche alle gleichzeitig in unterschiedlicher Stärke aktiv sind, sondern hier existiert eine Vielzahl von Komponenten, die nun aber eben nicht alle gleichzeitig aktiv sind. Diese Codierungsart steht in einem engem Zusammenhang zu assoziativen Speicherverfahren ist aber in der Informationstechnik nicht üblich *[Marr-82]*. Wegen der wesentlichen Bedeutung für biologische Nervennetze ist dieses Prinzip aktueller Forschungsgegenstand in den Neurowissenschaften.

Gerade in biologischen Systemen sollte natürlich die Maximalforderung eingehalten werden, dass die derart extrahierten Merkmale ein Signal möglichst kompakt darstellen und sich invariant gegenüber Störungen verhalten. Beide Forderung sind ja von der Signalcodierung bekannt, hier entfernt man zunächst die vorhandene Redundanz um eine kompakte Darstellung zu erhalten, um Störeinflüssen entgegenzuwirken und Fehler zu erkennen bzw. zu korrigieren wird anschließend Redundanz wieder hinzugefügt.

Ein wichtiges Gütekriterium für eine Codierung ist die Codeeffizienz ε. In *[Rieke-99]* wird bei der Definition dieser Größe auch die Reproduzierbarkeit des Originalsignals berücksichtigt. Für die Reproduktion muss dann natürlich auch immer die Decodierung realisiert werden. Für ein zu codierendes eindimensionales Signal X gilt bspw. nach (8.1.8):

$$\varepsilon = \frac{MI}{H(X)} = \frac{\frac{1}{2} ld(1+SNR)}{H(X)}$$

Das SNR ist dabei eine Funktion des Abstandes von Originalsignal und decodiertem Signal. Eine hohe Codeeffizienz erhält man demnach, wenn der Abstand zwischen diesen beiden Signalen möglichst gering ist.

Ein Hierarchisches Codiersystem (*HCS*) nach dem Vorbild von W. Hilberg bzw. J. Hawkins, berücksichtigt nun die beiden gerade genannten Aspekte. Unter Verwendung des *HCS* kann man demnach Paare bilden, die aus Original und Reproduktion bestehen. Für diese Paare können Abstände gebildet werden und auch die Codeeffizienz des Systems kann berechnet werden. Wenn man nun über eine Methode verfügt, mit der die Codeeffizienz optimiert werden kann, so liegt nach *[Hawkins-04]* eine kompakte und invariante Repräsentation von Signalen vor. Prinzipiell könnte man nun den *BPA* in das *HCS* integrieren, dann kann die auf das *HCS* verteilte Redundanz genutzt werden, um auch die Prozesse Erkennung/Synthese zu realisieren. D.h. aber, dass mit einem hierarchischen System das Potential vorliegt, Prozesse wie Codierung/Decodierung und Erkennung/Synthese einheitlich zu beschreiben. Die bisher mit jeweils unterschiedlichen Systemen realisierten Prozesse können demnach auf ein einziges System beschränkt bleiben. Sowohl die Speicherung als auch die Erkennung/Synthese erfolgt dann unter Verwendung der einmal gefundenen Codes.

Zu einer solchen Beschreibung sollte man gelangen können, wenn die Assoziationsmatrix in den *Belief-Propagation-Algorithmus* integriert wird. Dann werden nicht nur aufsteigende und absteigende Informationen zwischen den Ebenen miteinander verknüpft, sondern zusätzlich werden auch die syntaktischen Informationen innerhalb der einzelnen Ebenen berücksichtigt. Diese Sicht lässt folgende Interpretationen zu:

Deterministische Interpretation

- aufsteigende Informationen steuern die Codierung (Konvergenz)
- absteigende Informationen steuern die Decodierung (Divergenz)

Die deterministische Interpretation entspricht der Sprachmaschine nach W. Hilberg, hier werden die Assoziationsmatrizen für die horizontale Informationsverarbeitung innerhalb einer Netzwerkebene benötigt.

Statistische Interpretation:

- aufsteigende Information steuert die Erkennung
- absteigende Information steuert die Synthese

Die statistische Interpretation entspricht dem Gedächtnis-Vorhersage-Modell nach Hawkins, hier wird bei jedem Schichtübergang von Knoten zu Knoten eine *CPT*-Matrix für die vertikale Informationsverarbeitung benötigt.

Beide Interpretationen können nun verallgemeinert werden, so dass eine vereinheitliche Beschreibung von Codierung/Decodierung und Erkennung/Synthese möglich wird. Die folgenden Ausführungen versuchen den Rahmen für eine solche Vereinheitlichung abzustecken:

Codierung

Zunächst beginnt man mit der Codierung in ungestörter Umgebung, dabei können sowohl die Assoziationsmatrizen (implizite Prädiktion) als auch die *CPT*-Matrizen (explizite Prädiktion) trainiert werden. Primär werden bei der Codierung die Alphabete für jede einzelne Schicht gelernt, in jeder Ebene entsteht eine Netzwerkstruktur für die typischen Verknüpfungen der Symbole des jeweiligen Alphabets.

Um nun eine kombinatorische Explosion des Symbolvorrats in den höheren Schichten zu vermeiden, müssen Ähnlichkeitsbündel (Cluster) zur Verdichtung der Information gefunden werden. Sekundär werden die Assoziationsmatrizen und die *CPT*-Matrizen für jede Metaebene gelernt. Je weiter man in den Metaebenen nach oben steigt, desto weitreichendere Abhängigkeiten werden gelernt. Mit dem Training der Symbolvorräte in den einzelnen Ebenen lässt sich eine starke Redundanzreduktion erzielen, außerdem wird das syntaktische Sprachwissen (Grammatik) vollständig auf das hierarchische Netzwerk verteilt.

Decodierung

Bei der Decodierung müssen die Mehrdeutigkeiten, welche durch die Ähnlichkeitsbündelung entstanden sind, unter Verwendung der Assoziationsmatrix und der *CPT*-Matrix aufgelöst werden. Trotzdem wird die Decodierung immer zu grammatikalisch korrekten Schätzungen der ursprünglich codierten Symbolsequenzen führen. D.h. jeder ursprünglichen Symbolsequenz kann eine decodierte Symbolsequenz zugeordnet werden. Mit diesen Sequenzpaaren kann man nun die Codeeffizienz des Systems bestimmen. Mit der Codeeffizienz gewinnt man eine Aussage über die Güte des Systems. Bei genügend hoher Güte kann der Prozess Codierung/Decodierung abgeschlossen werden. Nach dessen Beendigung liegen dann sowohl die Assoziationsmatrizen für die implizite Prädiktion als auch die *CPT*-Matrizen für die explizite Prädiktion als Wissensquellen vor, diese können für die Erkennung bzw. die Synthese genutzt werden.

Erkennung

Für die Erkennung kann man nun neben der syntaktischen Information (implizite Prädiktion) sowohl die aufsteigenden Informationen als auch die absteigenden Informationen nutzen. D.h. neben der syntaktischen Information können nun auch die strukturellen Informationen eines hierarchischen Systems ausgenutzt werden. Das bedeutet zunächst, dass nach jeder Erkennung auf der obersten Schicht eine Vorhersage (implizite Prädiktion) erfolgen kann. Für die implizite Prädiktion werden allerdings Stützstellen benötigt.

Bei der Sprachmaschine wurden als Stützstellen die Leitworte verwendet, d.h. für Erkennungsaufgaben sollte man die als zuverlässig identifizierten Symbole verwenden (in der Spracherkennung z.B. Vokale). Ausgehend von der impliziten Vorhersage der obersten Hierarchieebene erfolgt nun eine absteigende Prädiktion (explizite Prädiktion) bis in die unterste Schicht, so dass für jede Schicht eine Erwartung vorliegt. Die vertikale Informationsverarbeitung oder explizite Prädiktion kann demnach auch als gleichzeitig wirksam werdende Synthese verstanden werden. Diesen Mechanismus würde man als Vorbahnung der Wahrnehmung interpretieren können.

Ein ähnliches Verhalten wurde bei den kürzlich entdeckten Spiegelneuronen beobachtet *[Rizolatti-06]* und wird von den Psychologen derzeit als eine Schlüsselfunktion für das Verstehen bzw. für das Erkennen der Bedeutung einer Handlung betrachtet *[Ramach.-99]*.

Unabhängig davon bleibt jedenfalls festzuhalten, dass bei der Erkennung in jeder Schicht prinzipiell auf prädiktive und diagnostische Informationen zurückgegriffen werden kann. Dabei kann ein hierarchisches System die absteigende Information der motorischen Ebene nutzen, um bei Konflikten die aufsteigende Information der sensorischen Ebene zu korrigieren.

Synthese

Bei der Synthese kehren sich die Bedeutungen von prädiktiver und diagnostischer Information lediglich um, denn hier benutzt das System die aufsteigende Information von der sensorischen Ebene zur Korrektur der absteigenden Information von der motorischen Ebene. Dieser Mechanismus ist in der Biokybernetik als Reafferenzprinzip bekannt.

Dabei wird zunächst eine Kopie vom aktuellen gedanklichen Symbol erstellt, anschließend wird dieses Symbol (die Hierarchie absteigend) für die Effektoren decodiert. Die sich tatsächlich einstellende Situation wird anschließend wieder afferent (die Hierarchie aufsteigend) codiert. Die Kopie des ursprünglichen gedanklichen Symbols (Efferenzkopie) kann nun mit dem zurückgeführten gedanklichen Symbol (Afferenz) verglichen werden. Die Differenz ist nun ggf. für eine Korrektur nutzbar, so dass mit dem weiteren Verlauf der Handlung das beabsichtigte Ziel erreicht werden kann.

10.4 Schlussbemerkung

In der vorliegenden Arbeit wurden die Kerneigenschaften des Peripheren Auditiven Systems (*PAS*) und des Zentralen Auditiven Systems (*CAS*) in einem robusten echtzeitfähigen Modell vereinigt. Nach Sichtung der einschlägigen Literatur konnten die folgenden Kerneigenschaften des *PAS* identifiziert werden: Gehörrichtige Filterbankanalyse, dynamische Komprimierung, Modulationsaspekte und Dekorrelation. Diese Eigenschaften wurden in unterschiedlicher Weise realisiert. Das Spektrum der Realisierungen reicht dabei vom klassischen *MFCC*-Modell über Hybride Modelle bis hin zu einem rein auditiven Modell. Neben den in der Literatur beschriebenen *VAMIG*-Ausprägungen, sind dabei auch neue Ansätze wie das auditive Waveletmodell oder die Implementierung der Lateralen Inhibition im Frequenz- und Zeitbereich unter Verwendung von dominanten Teilbandfrequenzen entwickelt worden.

Einen fairen Vergleich zwischen den verschiedenen *PAS*-Ausprägungen hinsichtlich deren Robustheit bei additiven Störungen ermöglichte die Einführung des Vereinheitlichten Auditiven Modells mit Integrierter Geräuschunterdrückung (*VAMIG*). Als Vergleichsmaß wurde die Phonartikulation nach H. Fletcher vorgeschlagen, mit diesem Maß kann die „Robustheit“ der einzelnen *VAMIG*-Ausprägungen quantifiziert werden. Es wurde gezeigt, dass dieses Robustheitsmaß eine Analogie zur Zuverlässigkeit von Parallelsystemen mit N-Elementen aufweist. Diese Analogie hat einerseits eine Konkretisierung des Begriffs Robustheit zur Folge, andererseits wurde die Messbarkeit der Robustheit ermöglicht. Untersuchungen unter Verwendung der Aurora-2 Datenbasis haben gezeigt, dass diejenigen *VAMIG*-Ausprägungen, welche die reinen auditiven Modelle repräsentieren, die robusteren *PAS*-Realisierungen darstellen. Da die Funktionalitäten Haarzelle und Laterale Inhibition nicht nur im Zeitbereich sondern auch im Frequenzbereich implementiert wurden, konnten reine auditive Modelle auch im Frequenzbereich implementiert werden. Somit lagen echtzeitfähige *PAS*-Modelle sowohl im Frequenz- als auch im Waveletbereich vor.

Diese *PAS*-Modelle stellen daher den Ausgangspunkt für die Einbindung zusätzlicher *CAS*-Eigenschaften dar. Im Gegensatz zu den *PAS*-Modellen sind Berichte über die Integration von *CAS*-Eigenschaften in echtzeitfähigen "embedded“ Systemen nach gegenwärtigem Kenntnisstand überhaupt noch nicht bekannt geworden. So bestand eine Herausforderung dieser Arbeit darin, echtzeitfähige Algorithmen zu finden, mit denen *CAS*-Eigenschaften realisiert werden können. Als Kerneigenschaften des *CAS* wurden dabei die Bindungseigenschaft und die Vorhersageeigenschaft identifiziert. Mit den in dieser Arbeit vorgeschlagenen Algorithmen konnte sowohl die Bindungseigenschaft (führte auf pitchkohärente Merkmale) als auch die Vorhersage zur Auflösung von Mehrdeutigkeiten (führte auf einen Evolutionär Orientierten Algorithmus) in einem echtzeitfähigen *PAS-CAS*-Modell integriert werden. Einschränkend musste jedoch festgestellt werden, dass diese Algorithmen nicht für jedes *PAS*-Modell geeignet sind. So konnte bspw. für das *PAS*-Modell unter Verwendung einer Wavelet-Zerlegung kein befriedigender Bindungsalgorithmus gefunden werden.

Die Leistungsfähigkeit der Merkmalsextraktionen von *PAS*-Modellen bzw. *PAS-CAS*-Modellen wurde für die Verbundworterkennung von Ziffern bei verschiedenen Störszenarien und unterschiedlichen Störintensitäten untersucht. Gemäß der Artikulationstheorie stellt diese Applikation sicher, dass die Erkennleistung nicht durch die Verwendung von Kontextwissen aus höheren Ebenen gesteigert werden kann. Unterschiede in der Leistungsfähigkeit können daher nur auf die Eigenschaften der verwendeten Modelle zurückgeführt werden. Darüber hinaus wurden im Training ausschließlich ungestörte Daten verwendet. Mit diesem Ansatz kann bei den anschließenden Erkenntests überprüft werden, wie zuverlässig die Modelle auf die große Bandbreite unterschiedlicher Störungen reagieren. Dabei konnte festgestellt werden, dass die *PAS-CAS*-Modelle bei additiven Störungen die besten Ergebnisse erzielen und sich daher „robuster“ verhalten als die einfachen *PAS*-Modelle.

Mit Tabelle 9.10 konnte diese Feststellung auch quantitativ untermauert werden. Gezeigt wurde dort der mittlere relative Gewinn der einzelnen *PAS-CAS*-Modelle gegenüber dem besten *PAS*-Modell, gemessen über allen Störgeräuschen und Störintensitäten. Dass dieser Gewinn nicht auf unterschiedliche Methoden der Geräuschunterdrückung zurückzuführen ist, wurde in Abschnitt 9, Tabelle 9.8 gezeigt.

Vor allem mit den hochauflösenden Varianten kann die Integration der beiden hypothetisch angegebenen *CAS*-Eigenschaften in einem *PAS-CAS*-Modell recht überzeugend gerechtfertigt werden. Hier konnten relative Verbesserungen gegenüber dem besten *PAS*-Modell von 17.42 % bzw. 18.34 % gemessen werden. Dies kann u.a. durch die stärkere Beachtung des Redundanz- und Kontinuitätsprinzips begründet werden. Einen ernsthaften Kritikpunkt an den *PAS-CAS*-Modellen stellt allerdings die Erkennleistung bei störungsfreien Szenarien dar, hier kann sicher noch weiteres Verbesserungspotential gefunden werden. Insgesamt konnte aber gezeigt werden, dass mit den vorgeschlagenen Algorithmen zur Einbindung von *CAS*-Eigenschaften deutliche Verbesserungen in den Erkennleistungen bei mittleren und schlechten SNR erzielt werden konnten.

Schließlich zeigen sich die *PAS-CAS*-Modelle auch dann noch ebenbürtig, wenn die Komponenten des *PAS*-Modells normalisiert werden und dass obwohl im Gegensatz zur Komponentennormalisierung (Normalisierung über der gesamten Äußerung) lediglich ein einzelner Frame der Äußerung berücksichtigt wird. Diese Beobachtung lässt den Schluss zu, dass auf der *CAS*-Ebene ein weiterer Mechanismus zur Selbstnormalisierung gefunden wurde.

Mit einem Ausblick auf weiteres Optimierungspotential und der Möglichkeit weitere strukturelle Eigenschaften des Neocortex in Modellen zur Merkmalsextraktion einzubinden, wurde darauf hingewiesen, dass die Bindungseigenschaft und insbesondere die Vorhersageeigenschaft ebenso in höheren Ebenen der Sprachverarbeitung anzutreffen sein wird. Es ist anzunehmen, dass in einem hierarchisch organisierten Sprachverarbeitungssystem weitere Steigerungen in den Erkennraten erreicht werden können.

Anhang A. Grundlagen der Statistik

A.1 Grundbegriffe

Für die Einführung des Begriffs der Wahrscheinlichkeit eines Ereignisses wird ein Zufallsexperiment betrachtet. Ein möglicher Versuchsausgang des Zufallsexperiments wird Ergebnis genannt und in der Statistik meist mit ω bezeichnet. Die Menge aller Ergebnisse wird Merkmalraum oder Ergebnisraum genannt und mit $\Omega = \{\omega\}$ bezeichnet. Ein Ereignis *A* stellt eine Menge (Zusammenfassung) von Ergebnissen dar. Stellen die Ergebnisse selbst Ereignisse dar, spricht man von so genannten Elementarereignissen ω_j. Die Wahrscheinlichkeit eines Ereignisses wird axiomatisch definiert, wobei die folgenden Eigenschaften gefordert werden (Kolmogoroff, 1933):

Axiom 1: $0 \le P(A) \le 1$

Axiom 2: $P[\Omega] = 1$

Axiom 3: $AB \neq 0 \Rightarrow P(A \cup B) = P[A] + P[B]$

Mit dem Merkmalraum Ω, der Menge aller Ereignisse A_i und der Wahrscheinlichkeitsfunktion $P[A_i]$ liegt ein vollständiges Wahrscheinlichkeitssystem $W=\{\Omega,\{A_i\},P[A_i]\}$ vor. Lässt sich ein Ereignis A als Vereinigung von paarweise disjunkten Ereignissen A_j darstellen,

$$A = A_1 \cup A_2 \cup \ldots \cup A_K = \bigcup_{i=1}^{K} A_i \text{ mit } \quad A_iA_j = \varnothing \text{ für } i \neq j,$$

dann gilt mit Axiom 3:

$$P[A] = \sum_{i=1}^{K} P[A_i] \tag{A.1.1}$$

Im Falle abzählbarer Elementarereignisse ω_j lässt sich jedes Ereignis A - unter Berücksichtigung der Disjunktheit von Elementarereignissen - aus einer Anzahl von Elementarereignissen zusammensetzen.

$$P[A] = \sum_{\omega_i \in A} P[\omega_i] \tag{A.1.2}$$

Für den gesamten Merkmalraum gilt dann entsprechend Axiom 2:

$$P[\Omega] = \sum_{alle\ \omega_i} P[\omega_i] = 1 \tag{A.1.3}$$

Für den Fall nichtabzählbarer Ergebnisse ω werden die Ergebnisse als Teil einer Zahlengerade betrachtet, z.B.

$$\Omega = \{\omega \in \Re \mid a \le \omega \le b\} \tag{A.1.4}$$

Jedes Ereignis A entspricht hier einer Vereinigung mehrerer zwischen a und b gelegener Intervalle. Die Wahrscheinlichkeit $P[A]$ eines Ereignisses wird dann gemäß (A.1.5) definiert

$$P[A] = \int_{|A} p(\omega) d\omega \tag{A.1.5}$$

Die Wahrscheinlichkeitsdichtefunktion $p(\omega)$ unterliegt entsprechend Axiom 3 der Einschränkung:

$$p(\omega) \geq 0 \quad \text{und} \qquad \int_{\Omega} p(\omega) d\omega = 1 \tag{A.1.6}$$

A.2 Zufallsvariablen und Verteilungsfunktionen

Die Versuchsausgänge von Zufallsexperimenten müssen nicht immer reellen Zahlen entsprechen. Trifft dieser Fall ein, kann das Konzept der Zufallsvariablen angewendet werden, den Ergebnissen ω eines solchen Zufallsexperiments werden dann reelle Zahlen *x* durch eine Abbildung $x=x(\omega)$ zugeordnet. Die Ergebnisse werden also auf Punkte der Zahlengerade abgebildet. Den verschiedenen Ereignissen entsprechen dann i.a. Vereinigungen von Intervallen bzw. isolierten Punkten der Zahlengeraden, z.B.

$$A = \{\omega \mid a \leq x(\omega) < b\}$$

$$B = \{\omega \mid x(w) = c\}$$

$$C = A \cup B = \{\omega \mid a \leq x(\omega) < b) \text{ oder } (x(\omega) = c)\}$$

Durch die Abbildung $x = x(\omega)$ erhalten wir eine zufällige Zahl *X*, die Zufallsvariable genannt wird. Bei einer Zufallsvariablen *X* entspricht also jedem Intervall der Zahlengeraden ein Ereignis A des zugrunde liegenden Zufallsexperiments. Die Wahrscheinlichkeit eines Intervalls der Zahlengeraden ist daher durch die Wahrscheinlichkeit *P[A]* des zugehörigen Ereignisses A gegeben.

$$P[A] = P[a \leq X < b] \qquad \text{für} \qquad A = \{w \mid a \leq x(\omega) < b\}$$

Die konkrete Berechnung der Wahrscheinlichkeit eines Intervalls stützt sich auf die Verteilungsfunktion $F_X(z)$ bzw. die Wahrscheinlichkeitsdichtefunktion $p_X(x)$.

Die Verteilungsfunktion $F_X(z)$ einer Zufallsvariablen *X* ist definiert als Wahrscheinlichkeit dafür, dass die Zufallsvariable *X* kleiner oder gleich der Zahl z ist $-\infty < X \leq z$.

$$F_X(z) = P[X \leq z] \tag{A.2.1}$$

Die Wahrscheinlichkeitsdichtefunktion $p_X(x)$ ist die Ableitung der Verteilungsfunktion:

$$p_X(x) = \frac{d}{dz} F_X(z) = \lim_{\Delta z \to 0} \frac{P[z < X \leq z + \Delta z]}{\Delta z} \tag{A.2.2}$$

Die Verteilungsfunktion ist eine nichtnegative, monoton von 0 auf 1 steigende Funktion, sie wird an der Stelle z durch das Integral

$$F_X(z) = \int_{-\infty}^{z} p_X(x) dx \tag{A.2.3}$$

berechnet. Die Wahrscheinlichkeit eines Intervalls $a < X \leq b$ berechnet sich mit zu:

$$P[a < X \leq b] = F_X(b) - F_X(a) = \int_{a}^{b} p_X(x) dx \tag{A.2.4}$$

Eine Zufallsvariable *X*, die nur diskrete Werte x_i mit der Wahrscheinlichkeit $P_i = P[X=x_i]$ annimmt, heißt diskret. Für die Verteilungsfunktion und die WDF einer diskreten Zufallsvariable erhält man:

$$F_{\mathbf{x}}(z) = \sum_i P_i u(z - x_i) \quad \text{bzw.} \qquad p_X(\xi) = \sum_i P_i \delta(z - x_i)\,, \qquad \text{(A.2.5)}$$

wobei es sich bei $u(z)$ und $\delta(z)$ um die Sprungfunktion bzw. den Dirac-Impuls handelt. Die Verteilungsfunktion ist demnach eine Treppenfunktion mit Sprunghöhen P_i an den Stellen x_i. Die WDF besteht aus Dirac-Impulsen mit Gewichten P_i an den Stellen x_i.

Darüber hinaus treten auch kombinierte Verteilungsfunktionen bzw. WDF von kontinuierlichen und diskreten Zufallsvariablen auf.

$$F_X(z) = F_c(z) + F_d(z) \qquad \text{und} \qquad p_X(x) = p_c(x) + p_d(x) \qquad \text{(A.2.6)}$$

Das folgende Beispiel illustriert eine solche kombinierte Zufallsvariable: Die kontinuierliche Zufallsvariable *X* wird durch Begrenzung in eine Zufallsvariable *Y* umgewandelt:

$$Y = \begin{cases} a, & X < a \\ X, & a \le X \le b \\ b, & X > b \end{cases} \qquad \text{(A.2.7)}$$

Für den kontinuierlichen Anteil der WDF von *Y* gilt zunächst:

$$p_c(y) = \begin{cases} 0, & y < a \\ p_X(y), & a < y < b \\ 0, & y < b \end{cases} \qquad \text{(A.2.8)}$$

Durch die Begrenzung kann *Y* die Werte < a bzw. > b nicht mehr annehmen und stimmt zwischen a und b mit *X* überein. Dagegen wird das gesamte Intervall $-\infty < x < a$ auf den Punkt *Y* = a abgebildet, weshalb die Ereignisse $-\infty < X < a$ und *Y* = a identisch sind und insbesondere gleiche Wahrscheinlichkeit haben:

$$P[Y = a] = P[-\infty < X < a] = \int_{-\infty}^{a} p_X(y)dy \qquad \text{(A.2.9)}$$

Das bedeutet, dass die WDF $p_Y(y)$ an der Stelle *Y* = a einen diskreten Anteil hat. Gleiches gilt für die Stelle *Y* = b:

$$P[Y = b] = P[b < X < \infty] = \int_{b}^{\infty} p_X(y)dy \qquad \text{(A.2.10)}$$

Der diskrete Anteil $p_d(y)$ der WDF kann mittels der Dirac-Funktion dargestellt werden:

$$p_Y(y) = P[Y = a]\delta(y - a) + P[Y = b]\delta(y - b) \qquad \text{(A.2.11)}$$

Die endgültige WDF wird entsprechend (A.2.6) durch $p_Y(y) = p_c(y) + p_d(y)$ beschrieben.

A.3 Erwartungswert und Varianz

Die wichtigsten Parameter für Wahrscheinlichkeitsverteilungen sind der Mittelwert und die Varianz, zunächst wird deren Definition für den eindimensionalen Fall angegeben:

$$E_X\{X\} = \mu = \int xp(x)\,dx \tag{A.3.0}$$

$$VAR\{X\} = \sigma^2 = E\{(X-\mu)^2\} \tag{A.3.1}$$

Die Wurzel der Standardabweichung wird mit σ bezeichnet. Die nächsten beiden Gleichungen geben die Berechnung von Mittelwert und Varianz für mehrdimensionale Zufallsvariablen an.

$$\boldsymbol{\mu} = E_X\{\mathbf{x}\} \tag{A.3.2}$$

$$\boldsymbol{\Sigma} = E_X\{(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T\} \tag{A.3.3}$$

A.4 Ausgewählte Dichtefunktionen

Eine besondere Rolle spielen Dichtefunktionen, welche durch Parameter beschrieben werden können. Zu ihnen zählen u.a. die Normalverteilung und die Gammaverteilung.

1-dimensionale Gaußverteilung:

$$p_X(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \tag{A.4.1}$$

d-dimensionale Gaußverteilung:

$$p_{\mathbf{x}}(\mathbf{x}) = \frac{1}{(2\pi)^{D/2}\,|\boldsymbol{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\} \tag{A.4.2}$$

Die Verteilungsfunktion der eindimensionalen Gaußverteilung der normalisierten Zufallsvariable $x = \frac{y-\mu}{\delta}$ wird nach (A.4.3) berechnet.

$$F_X(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\,dx \tag{A.4.3}$$

Das Integral der Verteilungsfunktion $F_X(z)$ einer gaußverteilten Zufallsgröße X ist nicht elementar lösbar. Einen Ausweg bietet die Überführung des Integrals in die Form der *erf*-Funktion, diese liegt für normierte Zufallsgrößen in Tabellenform vor.

$$erf(z) = \int_0^z \frac{2}{\sqrt{\pi}} e^{-t^2} dt \tag{A.4.4}$$

Berücksichtigt man, dass die WDF der Normalverteilung eine symmetrische Funktion bezüglich des Mittelwertes ist, kann die Lösung des Integrals in zwei Schritten erfolgen. Zunächst wird (A.4.3) zerlegt:

$$F_X(z) = \int_{-\infty}^{-z} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx + \int_{-z}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \tag{A.4.5}$$

bzw.

$$F_X(z) = \int_{-\infty}^{-z} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx + 2\int_{0}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \tag{A.4.6}$$

Im ersten Schritt wird der zweite Term in (A.4.6) kann mit der Substitution $\frac{x}{\sqrt{2}} = t$ und $\frac{dx}{dt} = \sqrt{2}$ in die Form der erf-Funktion gebracht werden, die Grenzen des Integrationsintervalls erstrecken sich dabei von t = 0 bis t = $\frac{z}{\sqrt{2}}$.

$$2\int_0^z \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = 2 \int_0^{t=\frac{z}{\sqrt{2}}} \frac{1}{\sqrt{\pi}} e^{-t^2} dt\sqrt{2} = \int_0^{\frac{z}{\sqrt{2}}} \frac{2}{\sqrt{\pi}} e^{-t^2} dt \tag{A.4.7}$$

D.h. das Integral über den mittleren Bereich der WDF berechnet sich zu:

$$\int_{-z}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = \int_0^{\frac{z}{\sqrt{2}}} \frac{2}{\sqrt{\pi}} e^{-t^2} dt = erf\left(\frac{z}{\sqrt{2}}\right) \tag{A.4.8}$$

Damit erhält (A.4.6) folgende Form:

$$F_X(z) = \int_{-\infty}^{-z} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx + erf\left(\frac{z}{\sqrt{2}}\right) \tag{A.4.9}$$

Der zweite Schritt berücksichtigt, dass die Normalverteilung eine symmetrische Verteilung ist, daher gilt für die Randbereiche der Wahrscheinlichkeitsdichtefunktion die Beziehung:

$$Q(z) = \int_{-\infty}^{-z} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = \int_z^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \tag{A.4.10}$$

Mit der Eigenschaft $F_X(z \to \infty) = 1$, folgt dann für (A.4.6):

$$1 = 2Q(z) + erf\left(\frac{z}{\sqrt{2}}\right) \tag{A.4.11}$$

bzw.

$$F_X(z) = Q(z) + erf\left(\frac{z}{\sqrt{2}}\right) \tag{A.4.12}$$

Die zwei letzten Gleichungen führen abschließend auf eine geschlossene Lösung für $F_X(z)$:

$$F_x(z) = \frac{1 + erf\left(\frac{z}{\sqrt{2}}\right)}{2} \tag{A.4.13}$$

A.5 Entropie und Transinformation

Die Entropie einer Zufallsvariable *X* mit der Verteilungsdichte *p(x)* ist definiert durch

$$H(X) = -\sum p(x) \log_2 p(x) \tag{A.5.1}$$

und misst die mittlere Unsicherheit der Zufallsvariable in Bits. Die bedingte Entropie *H(X|Y)* ist Entropie der Zufallsvariablen *X* bei Vorgabe einer weiteren Zufallsvariable *Y*. Die Reduktion der Unsicherheit aufgrund einer gegebenen Zufallsvariable *Y* wird Transinformation (*engl. Mutual Information*) genannt.

$$I(X;Y) = H(X) - H(X \mid Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)\,p(y)} \tag{A.5.2}$$

Mit der Transinformation kann die Abhängigkeit zwischen zwei Zufallsvariablen gemessen werden. Die Transinformation ist ein Spezialfall der sogenannten Kullback-Leibler-Distanz, sie wird oft auch als relative Entropie oder Kreuzentropie bezeichnet.

$$D(p \parallel q) = \sum_x p(x) \log \frac{p(x)}{q(x)} \tag{A.5.3}$$

Diese Distanz kann als ein Maß für den Abstand von zwei Verteilungsdichten *p(x)* und *q(x)* betrachtet werden.

A.6 Modell des diskreten Übertragungskanals

In der Informationstheorie wird das folgende Modell einer aus Sender, Kanal und Empfänger bestehenden Übertragungsstrecke betrachtet. Der Sender hat eine Menge $S_S=\{s_1,s_2,\ldots,s_M\}$ von diskreten Symbolen zur Verfügung, aus denen er ein Symbol s_i mit der Wahrscheinlichkeit $P(s_i)$ auswählt. Der Vorgang „Senden" ist demnach ein Zufallsexperiment, bei dem aus einer Menge S_s beliebige Elementarereignisse s_i ausgewählt werden. Der Empfänger empfängt nun Symbole s_j aus der Menge $S_E=\{s_1,s_2,\ldots,s_R\}$. Bei einem gestörten Kanal kann der Vorgang „Empfangen" ebenfalls statistisch beschrieben werden.

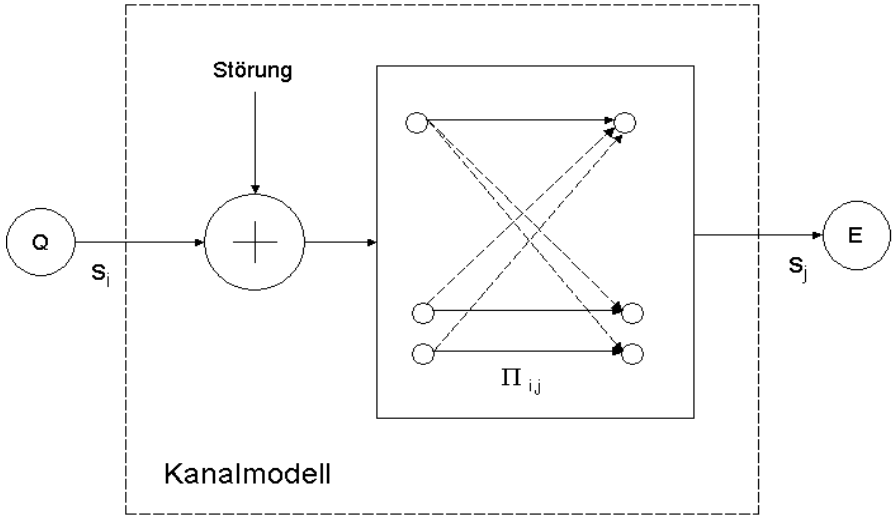

Abbildung A.6.1: ***Shannonsches Kanalmodell, der akustische Kanal wird informationstheoretisch als rauschender Kanal verstanden, über den M-Symbole übertragen werden. Die gesendeten Symbole s_i besitzen eine a-priori Wahrscheinlichkeit und werden gemäß der Übergangsmatrix Π in die Symbole s_j transformiert.***

Der Kanal bewirkt eine Abbildung $s_i \rightarrow s_j$. Wenn es sich um einen gestörten Kanal handelt, kann diese Abbildung statistisch durch die Angabe von Übergangswahrscheinlichkeiten beschrieben werden. Zweckmäßigerweise fasst man alle Übergangswahrscheinlichkeiten $P(s_i \mid s_j)$ zu einer Matrix Π zusammen. Die gesamte Übertragung ist somit durch die Menge der Symbole S_S und S_E, die Sendewahrscheinlichkeit $P(s_i)$ und die Übergangsmatrix Π vollständig beschrieben. Die Sendewahrscheinlichkeit des Symbols s_i berechnet sich über die relative Häufigkeit der gesendeten Symbole:

$$P(s_i)=\frac{N_i}{N} \qquad N=\sum_{i=1}^{M} N_i \tag{A.6.1}$$

Die Elemente der Übergangsmatrix $\Pi_{i,j}$,also die Übergangswahrscheinlichkeiten $P(s_j \mid s_i)$, werden nach (A.6.2) geschätzt.

$$\Pi_{i,j} = \frac{\frac{N_{i,j}}{N}}{\frac{N_i}{N}} = \frac{N_{i,j}}{N_i} \qquad \text{(A.6.2)}$$

$N_{i,j}$ entspricht der absoluten Häufigkeit für das gemeinsame Auftreten der Symbole s_i und s_j. Die absolute Häufigkeit des Symbols *i* wird mit N_i bezeichnet, die Zahl *N* entspricht der Gesamtanzahl gesendeter Symbole. Die Berechnung der Wahrscheinlichkeit, dass das Symbol s_i gesendet wurde wenn das Symbol s_j empfangen wurde, erfolgt gemäß der Bayes'schen Beziehung:

$$P(s_i \mid s_j) = \frac{P(s_j \mid s_i)P(s_i)}{P(s_j)} \qquad \text{(A.6.3)}$$

Die Wahrscheinlichkeit $P(s_i \mid s_j)$ bezeichnet man diesem Zusammenhang oft als a-posteriori Wahrscheinlichkeit, denn sie steht für den Zuwachs an Wissen, wenn man zusätzlich zur a-priori Wahrscheinlichkeit $P(s_i)$ auch noch die sogenannte Likelihood $P(s_j \mid s_i)$ beobachten kann.

Anhang B. Die Artikulationstheorie

„How do Humans Recognize and Process Speech“ J. B. Allen, 1994

Jont B. Allens Artikel mit dem oben genannten Titel hat wesentlich zur Motivation der vorliegenden Arbeit beigetragen. Der Artikel stellt eine Zusammenfassung und Interpretation der Arbeiten von Harvey Fletcher dar. H. Fletcher und seine Kollegen von den AT&T Bell Labs arbeiteten von 1918 –1950 an der quantitativen Bestimmung der Sprachqualität über gestörte Telefonverbindungen und der Verbesserung der Verständlichkeit (*engl.* Intelligibility). Als Ergebnis dieser Arbeiten entstand ein fünfschichtiges Modell der menschlichen Sprachverarbeitung. Aus heutiger Sicht scheinen die Ergebnisse dieser Experimente auch für den Entwurf von *ASR*-Systemen relevant zu sein. Eine Vielzahl von Veröffentlichungen sind von J.B. Allens Artikel inspiriert worden. Um diesem Umstand Rechnung zu tragen, sollen in diesem Abschnitt die wesentlichen Ergebnisse zusammengefasst werden.

B.1 Kanalmodell, Kontext und Entropie

H. Fletcher fasste Sprache als eine zeitliche Folge von sprachlichen Einheiten welche bei der Übertragung über einen Kanal gestört werden können. Mit Verwendung des Kanalmodells nutzte Fletcher intuitiv wesentliche Ergebnisse der erst später von Shannon veröffentlichten Informationstheorie.

Für seine Experimente verwendete H. Fletcher so genannte sinnlose Silben. Als Symbole eines Alphabets wurden Kombinationen von CVC-Silben (Konsonant-Vokal-Konsonant) und Kombinationen von CV- bzw. VC-Silben verwendet. Das Konzept der Entropie spielt bei Fletchers Überlegungen zur Auswahl der Symbole eine große Rolle. So kann eine Folge von sprachlichen Einheiten auf unterschiedlichen Ebenen (Phonem, Silbe, Wort und Satz) betrachtet werden. Diese Ebenen unterscheiden sich vor allem in der Stärke des vorhandenen Kontext. Kontext spielt bei der Erkennung von Sprache eine wesentliche Rolle. Als Beispiel wurden in *[Allen-94]* die Sätze “How do humans recognize speech“ und “How do humans wreck a nice beach“ angegeben, deren Unterscheidung nur im Kontext möglich ist. Zur quantitativen Beschreibung von Kontext ist die Entropie geeignet: so ist die Entropie H_{CVC} von sinnlosen Silben bspw. größer als die Entropie H_w von Worten oder Sätzen H_S.

$$H_{CVC} > H_w > H_S \tag{B.1.1}$$

Nimmt man für eine Sprache etwa 40 mögliche Phoneme an, erhält man für CVC-Silben bei Gleichverteilung der Phoneme eine maximale Entropie von 3log(40) ≈ 16 Bit. Worte dagegen, zeichnen sich im Gegensatz zu sinnlosen Silben durch eine Bedeutung aus, damit verbunden ist das Vorhandensein von Kontext bzw. höherem Wissen. Dieses höhere Wissen setzt der Mensch bei der Dekodierung der akustischen Signale ein. Mit dem Konzept der Entropie kann also erklärt werden, warum bei gleicher Kanalstörung die Erkennraten für Worte bzw. Sätze höher sind als bei einem Vokabular, welches sich ausschließlich aus sinnlosen Silben zusammensetzt. Um nun die grundlegenden Mechanismen der menschlichen Spracherkennung zu untersuchen, muss die Verwendung von höherem Wissen ausgeschlossen werden, daher die Verwendung eines Corpus mit sinnlosen Silben.

B.2 Das Artikulationsexperiment

Nachdem eine feste Datenbasis mit statistisch ausbalancierten sinnlosen *CVC-, CV-* und *VC-* Silben und konstanter Quellenentropie vorlag, wurden deren Elemente durch verschiedene Sprecher über gestörte Telefonkanäle übertragen. Den Hörern wurde die Aufgabe gestellt, aus einer geschlossenen Menge von möglichen sinnlosen Silben die gehörten Silben auszuwählen. Mit diesen Ergebnissen konnten einige empirische Wahrscheinlichkeiten gemessen werden. H. Fletcher verwendete hierbei folgende Definitionen:

Articulation: relative Häufigkeit korrekter Erkennung, wenn kein Kontext vorliegt

(Erkennung von sinnlosen Silben, Konsonanten oder Vokalen)

Intelligibility: relative Häufigkeit korrekter Erkennung, wenn Kontext vorliegt

(Erkennung von Worten oder Sätzen)

Außerdem konnte die Artikulation durch Änderung des SNR und komplementäre Hochpass-Tiefpassfilterung mit unterschiedlicher Grenzfrequenz variiert werden. Darüber hinaus war Fletcher klar, dass die Entropie einer Quelle welche voneinander unabhängige Symbole sendet größer ist, als die Entropie einer Quelle welche voneinander abhängige Symbole sendet.

$$H_{Bed} < H_Q \tag{B.2.1}$$

Eine Besonderheit liegt darin, dass sowohl die gesendeten als auch die empfangenen Symbole aus dem gleichen Merkmalraum stammen. Gemäß A.1 werden die Sendewahrscheinlichkeiten $P(s_i)$ und die Übergangsmatrix $\Pi(\alpha)$ gemessen. Der Faktor α wird zur Verstärkung des Sprachsignals verwendet, um verschiedene *SNR* realisieren zu können. Für die Matrix lassen sich dann zwei Grenzfälle ableiten:

($\alpha=1$, *SNR*≈*30dB*), d.h. die Übertragungsbedingungen sind ideal, die Matrix $\Pi(\alpha)$ liegt in Diagonalform vor:

$$P(s_j \mid s_i) = \delta_{i,j} = \begin{cases} 1 \ , & \text{für } i = j \\ 0 \ , & \text{sonst} \end{cases} \tag{B.2.2}$$

Für die Summe der Diagonalen gilt demnach:

$$sp\ \Pi(\alpha = 1) = M \tag{B.2.3}$$

($\alpha=0$, *SNR*≈*0dB*), d.h. bei schlechtem *SNR* haben alle Elemente $\Pi_{i,j}(\alpha)$ etwa die gleiche Wahrscheinlichkeit.

$$P(s_j \mid s_i) = \frac{1}{M} \quad , \text{für alle } i, j = 1..M \tag{B.2.4}$$

Hier ergibt sich für die Summe über alle Diagonalelemente:

$$sp\ \Pi(\alpha = 0) = 1 \tag{B.2.5}$$

Um die Berechnungen für die verschiedenen Datenbasen {C,V,C} und {C,V} zu vereinheitlichen, wurde zusätzlich eine mittlere Phonartikulation berechnet:

$$s(\alpha) = (2c + v)/3 \tag{B.2.6}$$

Da sich die Phonartikulation gemäß der Definition von H. Fletcher nur auf die korrekt erkannten Phone bezieht, folgt schließlich:

$$s(\alpha) = \frac{1}{M}\sum_{i=1}^{M}\Pi_{i,i} \tag{B.2.7}$$

Für den Artikulationsfehler folgt mit $e(\alpha) = 1 - s(\alpha)$, also bei Summation über alle Nichtdiagonalelemente:

$$e(\alpha) = \frac{1}{M}\sum_{i=1}^{M}\sum_{j'=1}^{M}\Pi_{i,j}(\alpha) \tag{B.2.8}$$

Fletcher fand heraus, dass die Silbenartikulation *S(α)* – also die Wahrscheinlichkeit der korrekten Erkennung von *CVC*-Silben – einerseits durch die Phonartikulation *c(α)* und *v(α)* korrekt vorhergesagt werden kann und andererseits durch die dritte Potenz der mittleren Phonartikulation sehr gut angenähert werden kann.

$$S(\alpha) = c(\alpha)v(\alpha)c(\alpha) \approx s^3(\alpha) \tag{B.2.9}$$

Dieses Ergebnis besagt, dass die drei Lauteinheiten bzw. eine sinnlose Silbe als unabhängige Lauteinheiten gehört werden. Um eine Silbe korrekt zu erkennen, müssen demnach alle drei Lauteinheiten nacheinander korrekt erkannt werden. Auf Kontextwissen kann auf dieser Ebene nicht zurückgegriffen werden. D.h. das Problem robuster HSR kann in zwei Teilprobleme zerlegt werden: Decodierung der Phone und Verwendung von Kontextwissen (Entropie) um Korrekturen vorzunehmen bzw. fehlende Informationen einzufügen.

Nachdem die fundamentale Bedeutung der Phonartikulation gezeigt werden konnte, untersuchten Fletcher und seine Kollegen die Phonartikulation in unterschiedlichen Frequenzbändern für verschiedene Grenzfrequenzen f_c einer Hochpass-Tiefpassweiche. Zunächst stellten sie fest, dass es sich bei den Teilbandartikulationen $s_L(f_c,\alpha)$ und $s_H(f_c,\alpha)$ erwartungsgemäß um monoton steigende bzw. monoton fallende Funktionen handelt, die sich aber nicht zur Gesamtbandartikulation addierten. Daher suchten sie nach einer Transformation welche die folgende Forderung erfüllt.

$$A(s_L(f_c,\alpha)) + A(s_H(f_c,\alpha)) = A(s(\alpha)) \tag{B.2.10}$$

H. Fletcher fand diese Transformation empirisch und nannte sie Artikulationsindex:

$$A(s) = \frac{\log_{10}(1-s)}{\log_{10}(1-s_{\max})} \tag{B.2.11}$$

Für perfekte Bedingungen (α → 1, keine HP/TP – Filterung) erhält man für s_{max} den Wert 0.985. Löst man die Gleichung nach s(A) auf erhält man:

$$s(A) = 1 - e_{\min}^{A} \tag{B.2.12}$$

Die nichtlineare Transformation *A(s)* transformiert $s(f_c,\alpha)$ in ein Integral über der Artikulationsindexdichte *D(f)*.

Dies wird ersichtlich, wenn jedem Term in (B.2.10) ein Integral *D(f)* über der Frequenz *f* zugeordnet wird.

$$A(s_L(f_c)) = \int_0^{f_c} D(f)df \qquad \text{(B.2.13)}$$

$$A(s_H(f_c)) = \int_{f_c}^{\infty} D(f)df \qquad \text{(B.2.14)}$$

$$A(s) = \int_0^{\infty} D(f)df \qquad \text{(B.2.15)}$$

Die Dichtefunktion ist folgendermaßen definiert:

$$D(f_c) = \frac{\partial}{\partial f_c} A(s_L(f_c)) \qquad \text{B.2.16)}$$

Aus den oben genannten Beziehungen lässt sich durch einfache Umformungen ein Kanalmodell mit unabhängigen Teilbändern entwickeln:

$$\frac{\log_{10}(1-s_L)}{\log_{10}(1-s_{\max})} + \frac{\log_{10}(1-s_H)}{\log_{10}(1-s_{\max})} = \frac{\log_{10}(1-s)}{\log_{10}(1-s_{\max})}$$

$$\rightarrow \quad \log_{10}[(1-s_L)(1-s_H)] = \log_{10}(1-s)$$

$$\rightarrow \quad (1-s_L)(1-s_H) = (1-s) = e_L e_H = e \qquad \text{(B.2.17)}$$

oder: $$s = s_L + s_H - s_L s_H \qquad \text{(B.2.18)}$$

D.h. die Teilbandartikulationen addieren sich nur für den Fall: $s_L s_H = 0$, also nur für disjunkte Merkmalsräume. Nach (B.2.17) sind die Artikulationsfehler im Tiefpasskanal unabhängig vom Hochpasskanal, demnach werden partielle Merkmale eines Sprachlauts unabhängig voneinander verarbeitet. Dieses Konzept lässt sich auf *K*-Kanäle zum so genannten Multiband-Modell erweitern:

$$e = e_1 e_2 \ldots e_K \quad \text{bzw.} \quad s = 1 - \prod_{k=1}^{K} e_k \qquad \text{(B.2.19)}$$

Fletcher zeigte hier, dass Phone in unabhängigen Artikulationsbändern verarbeitet werden. Auffällig an der Struktur von (B.2.19) ist, dass der gesamte Artikulationsfehler nie größer werden kann als der kleinste Teilbandartikulationsfehler. Interessanterweise werden Strukturen dieser Art in der Technik oft verwendet, um die Zuverlässigkeit von Systemen zu erhöhen. Tatsächlich wurden vor diesem Hintergrund Verfahren entwickelt, mit denen die Zuverlässigkeit aller Teilbänder berechnet wurde und zur Erkennung lediglich die zuverlässigsten Bänder benutzt wurden.

Kurze Zeit später wurde noch ein weiterer wichtiger Zusammenhang gefunden:

$$e_k = e^{D_k \min} \tag{B.2.20}$$

mit

$$D_k = \int_{f_k}^{f_{k+1}} D(f)\,df \tag{B.2.21}$$

Die Frequenzgrenzen werden so gewählt, dass alle *k*-Teilflächen unter *D(f)* gleich groß sind. Fletcher fand nun heraus, dass die Artikulation D_k als Funktion von *α* lediglich vom *SNR(dB)* in diesem Teilband abhängt.

$$D_k(\alpha) = \frac{1}{k} SNR_k(\alpha)/30 \tag{B.2.22}$$

Für α → 1 erhält man demnach mit (B.3.2) $e = e_{\min}$ bzw. $s_{\max} = 1 - e_{\min}$. Mit dieser Beziehung wird deutlich, dass das SNR und nicht die Energie die Phonartikulation *s* bestimmt.

B.3 Das Artikulationsmodell

Nun konnte H. Fletcher ein Modell aufbauen mit dem es möglich war, ausgehend von den Teilband *SNR* die Phonartikulation, die Silbenartikulation und die Wortartikulation vorherzusagen. Dieses Modell wurde für eine Vielzahl von Kombinationen unterschiedlicher Kanalparameter gestestet und zeigte eine beeindruckende Genauigkeit über einen weiten Bereich unterschiedlicher Kanalbedingungen.

$$A(\alpha) = \sum_{k=1}^{K} D_k(\alpha) \tag{B.3.1}$$

$$s(A) = 1 - \prod_{k}^{K} e_{\min}^{D_k(\alpha)} = 1 - e_{\min}^{A(\alpha)} \tag{B.3.2}$$

$$S(A) = s^3 \tag{B.3.4}$$

$$W(A) = 1 - (1 - S(A))^j \tag{B.3.4}$$

Der Parameter *j>1*, hängt von der Entropie der Datenbasis ab und wird empirisch bestimmt. Mit diesem Modell und insbesondere der Bestimmung der Teilbandartikulation D_k erhält man Hinweise auf Möglichkeiten zur Erhöhung der Zuverlässigkeit der Phonartikulation.

Interpretation: Robuste *ASR*-Systeme sollten sich demnach durch den Einsatz von solchen Algorithmen auszeichnen, welche der Verbesserung der Teilband-*SNR* dienen. Offensichtlich sind diejenigen Gebiete in der Zeit- und Frequenzebene von Bedeutung, in denen das lokale *SNR* besonders gute Werte erreicht. Hinsichtlich einer effizienten Informationsübertragung sollten die wichtigsten linguistischen Informationen durch besonders robuste Signaleigenschaften repräsentiert werden. Dies gilt bspw. für Formanten, hier ist das lokale *SNR* i.d.R. am größten. Darüber hinaus lässt die Analyse von auditiven Modellen vermuten, dass biologisch motivierte Ansätze, namentlich Laterale Inhibition und Adaption ebenfalls dem Zweck dienen, das lokale *SNR* zu erhöhen.

Während die Laterale Inhibition zu einer Schärfung der spektralen Spitzen führt, zeigen Haarzellenmodelle dank ihrer dynamischen Eigenschaften ebenfalls einen Trend zur lokalen Verbesserung des *SNR*: Schnelle Änderungen im Signal werden verstärkt, während dauerhafte Erregung allmählich gedämpft werden.

Nach einer derartigen Vorverarbeitung werden voneinander unabhängige lokale Merkmale in den Teilbändern extrahiert und unter Berücksichtigung der Zugehörigkeit zu einem auditiven Objekt zusammengesetzt und höheren Verarbeitungsstufen zugeführt. Hier findet also so etwas wie eine Überführung in Symbolfolgen statt, so dass in den nachfolgenden Stufen zunehmend Kontextinformation verwendet werden kann, um Inkonsistenzen in den Symbolfolgen zu identifizieren und zu korrigieren.

Anhang C. Wavelets und Filterbänke

C.1 Orthogonale Funktionensysteme

Eine wichtige Anwendung von orthogonalen Funktionensystemen ist die redundanzfreie und kompakte Approximation von Signalen.

$$x(t) = \sum_m c_m \psi_m \qquad m \in M \tag{C.1.1}$$

Dabei stellen die ψ_m eine Menge von orthogonalen Basisfunktionen dar. Die Berechnung der Koeffizienten c_m gestaltet sich besonders einfach, wenn die ψ_m zu einem orthonormalen Funktionensystem gehören.

$$< \psi_m, \psi_{m'} > = \delta_{m,m'} \tag{C.1.2}$$

In diesem Fall erhält man die Koeffizienten durch Berechnung des Skalarproduktes.

$$c_m = < \psi_m, x(t) > \tag{C.1.3}$$

Die Fourierreihe von periodischen Funktionen ist ein klassisches Beispiel. Hier werden komplexe periodische Basisfunktionen im Intervall $\infty < t < \infty$ verwendet.

$$x(t) = \sum_m c_m e^{j\omega m t} \tag{C.1.4}$$

Für stationäre Signale erhält man mit diesem Ansatz gute Approximationen, hingegen sind lokal begrenzte Basisfunktionen bei nichtstationären Signalen die bessere Wahl. Zur Klasse der nicht-stationären Signale gehören auch die Sprachsignale. Neben den periodischen und aperiodischen Abschnitten wechseln sich hier energiereiche Signalsegmente mit stillen Abschnitten ab. Hier wird bereits eine Problematik sichtbar. Die verschiedenen Basisfunktionen sollen sich in bestimmten Signalabschnitten auslöschen in anderen hingegen verstärken. Diesen abrupten Wechsel zwischen Auslöschung und Verstärkung kann man nicht mit wenigen Basisfunktionen erreichen. D.h. für eine gute Approximation müssen offenbar sehr viele Basisfunktionen verwendet werden. Außerdem führt das Weglassen der Koeffizienten c_m zu einem Fehler in der Darstellung von *x(t)*, der sich über den ganzen Definitionsbereich erstreckt. Besonders stark wirken sich Unstetigkeiten von *x(t)* aus, die Rekonstruktion des Signals zeigt dann Welligkeiten innerhalb des gesamten Definitionsbereichs.

C.2 Kurzzeit-Fouriertransformation (STFT)

Bessere lokale Eigenschaften erhält man bei Verwendung von Fensterfunktionen *g(t)* begrenzter Ausdehnung. Berechnet man die Fouriertransformation des durch die Fensterfunktion an der Stelle τ ausgeblendeten Signalausschnitts, erhält man die Kurzzeit-Fouriertransformation

$$STFT(\tau, \omega) = \int x(t) g(t-\tau) e^{-j\omega t} dt \tag{C.2.1}$$

Mit (C.2.1) liegen zwei Interpretationsmöglichkeiten vor. Entweder berechnet man alle auftretenden Frequenzkomponenten zum Zeitpunkt τ , oder aber man berechnet den Verlauf einer Frequenzkomponente ω über der Zeit.

Der letztere Fall entspricht formal einer Bandpassfilterung, wobei man die jeweilige Impulsantwort des Bandpassfilters durch Modulation von *g(t)* mit der komplexen Exponentialfunktion der Frequenz ω erhält. Die *STFT* lässt sich demnach auch als Filterbank interpretieren. Grundsätzlich wird die *STFT* als Abbildung in die Zeit- Frequenz Ebene verstanden. Die Eigenschaften der *STFT* hängen aber offensichtlich von der verwendeten Fensterfunktion *g(t)* ab. Sowohl für die Fouriertransformierte *G(f)* als auch für die Zeitfunktion *g(t)* des Fensters kann deren lokale Ausdehnung angegeben werden.

$$\Delta f^2 = \frac{\int f^2 |G(f)|^2 df}{\int |G(f)|^2 df} \qquad \text{bzw.} \qquad \Delta t^2 = \frac{\int t^2 |g(t)|^2 dt}{\int |g(t)|^2 dt} \qquad \text{(C.2.2)}$$

In beiden Fällen entspricht der Nenner in den obigen Ausdrücken jeweils der Energie der verwendeten Fensterfunktion. (C.2.2) zeigt, dass zwei sinusförmige Frequenzen getrennt werden können, wenn sie sich um mindestens Δf unterscheiden. Ähnliches gilt für die Unterscheidung von 2 aufeinander folgenden Impulsen, nur wenn die beiden Impulse um mindestens Δt auseinander liegen ist eine eindeutige Trennung möglich. D.h. die Auflösung der *STFT* ist durch Δf bzw. Δt gegeben. Nach der Heisenbergschen Ungleichung (C.2.3) kann die Auflösung von Frequenz und Zeit nicht beliebig klein werden.

$$\Delta f \, \Delta t \geq \frac{1}{4\pi} \qquad \text{(C.2.3)}$$

Das bedeutet, man kann entweder die Zeit- oder die Frequenzauflösung verbessern. Noch wichtiger ist aber die folgende Feststellung: Hat man sich zur Berechnung der *STFT* für eine Fensterfunktion entschieden, so ist auch die Bandbreite der Filter und somit die Frequenz- und Zeitauflösung für die gesamte Zeit- Frequenzebene festgelegt. Nun wäre es aber wünschenswert, die Auflösung Δf und Δt innerhalb der Zeit-Frequenzebene variabel zu gestalten. Bspw. sollten die tiefen Frequenzen eines Signals mit guter Frequenzauflösung und die hohen Frequenzen mit guter zeitlicher Auflösung analysiert werden können. Eine solche Analyse mit Mehrfachauflösung liefert die Wavelet-Transformation.

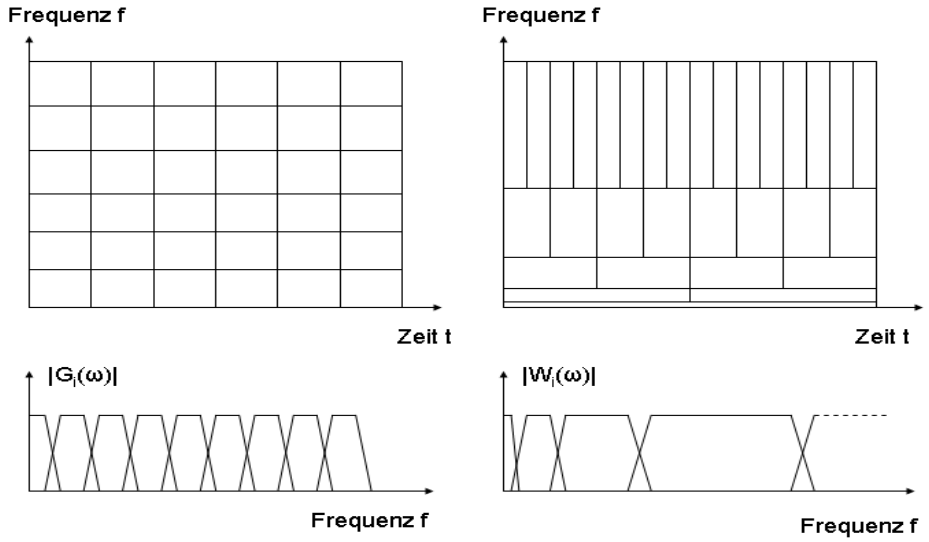

Abbildung C.1: ***Im linken Teil wird die Zeit-Frequenzauflösung für die STFT gezeigt, der rechte Teil der Abbildung zeigt eine ungleichmäßige Verteilung der Zeit- Frequenzauflösung oder auch Mehrfachauflösung, so wie sie bei der Signalzerlegung mit einer WT anzutreffen ist. Im unteren Teil der Abbildung sind qualitativ die Beträge der Frequenzgänge der zugehörigen Filterbänke dargestellt. Die Bandbreite der Filter ist bei der STFT konstant, bei der WT dagegen ist die relative Bandbreite $\Delta f/f_m$ konstant, die Mittenfrequenz f_m verdoppelt sich von Filter zu Filter.***

C.3 Mehrfachauflösung, Wavelet-Transformation (WT)

Es existieren ineinander verschachtelte Signalunterräume des Signalraums $L_2(\Re)$ mit unterschiedlicher Auflösung.

$$V_{-\infty} \subset ... \subset V_{-2} \subset V_{-1} \subset V_0 \subset V_1 \subset V_2 \subset ... \subset V_{\infty} \tag{C.3.1}$$

Vollständigkeit und Abgeschlossenheit:

$$\bigcap_{j \in Z} V_j = \{0\} \qquad \bigcup_{j \in Z} V_j = L_2(\Re) \tag{C.3.2}$$

Es existiert eine sogenannte Skalierungsfunktion $\Phi(t) \in V_0$, so dass $\forall\, j, k \in Z$ die Menge der

$$\{\Phi_{j,k}(t) = 2^{j/2}\Phi(2^j t - k)\} \tag{C.3.3}$$

eine orthonormale Basis des Raumes V_j darstellen.

$$< \Phi_{j,k}(t), \Phi_{j,k'}(t) > = \delta_{k-k'} \tag{C.3.4}$$

D.h. die Skalierungsfunktion spannt zusammen mit ihren Translationen den Raum V_j orthogonal auf. Benachbarte Unterräume sind über die Skalierungseigenschaft miteinander verknüpft:

$$f(x) \in V_j \Leftrightarrow f(2x) \in V_{j+1} \tag{C.3.5}$$

Darüber hinaus lässt sich zu jedem Unterraum $V_j \subset V_{j+1}$ ein orthogonales Komplement W_j definieren, so dass der Raum V_{j+1} als direkte Summe dargestellt werden kann.

$$V_{j+1} = V_j \oplus W_j \tag{C.3.6}$$

W_j entspricht dem Differenzraum zweier aufeinander folgender Skalierungsräume und wird durch eine orthogonale Basis von verschobenen Waveletfunktionen $\Psi(t)$ aufgespannt.

$$\Psi_{j,k}(t) = 2^{j/2}\Psi(2^j t - k) \tag{C.3.7}$$

Es ist zu beachten, dass die Waveletfunktionen $\Psi_{j.k}(t)$ einerseits auf allen Skalen orthonormal zueinander sind

$$< \Psi_{j,k}(t), \Psi_{J,K}(t) > = \delta_{k-K}\delta_{j-J} \tag{C.3.8}$$

und andererseits zu den Skalierungsfunktionen Orthogonalität vorliegt.

$$< \Phi_{j,k}(t), \Psi_{j',k'}(t) > = 0 \tag{C.3.9}$$

Mit (C.3.1) und (C.3.6) folgt für den in der Praxis auftretenden Fall die Beziehung

$$V_{J+1} = V_0 + \sum_{j=0}^{J} W_j \tag{C.3.10}$$

D.h. die Zerlegung setzt sich aus einer Komponente des Raumes V_0 mit der gröbsten Auflösung und mit weiteren Komponenten aus Räumen mit feinerer Auflösung zusammen. Die Basisfunktionen setzen sich demnach aus der Skalierungsfunktion der Stufe *j=0* und den Waveletfunktionen der Stufen $0 \le j \le J$ zusammen.

C.4 Die Diskrete Wavelet-Transformation (DWT)

Der Skalierungseigenschaft (C.3.5) kann man entnehmen, dass sich die Skalierungsfunktion $\Phi(t) \in V_0$ mit (C.3.1) oder genauer mit $V_0 \subset V_1$ als Linearkombination von Skalierungsfunktionen aus dem Unterraum V_1 berechnen lässt.

$$\Phi(t) = \sqrt{2} \sum_k c(k) \Phi(2t-k) \quad \text{mit} \quad \Phi(2t-k) \in V_1 \tag{C.4.1}$$

Für die Waveletfunktion $\Psi(t) \in W_0$ gilt analog:

$$\Psi(t) = \sqrt{2} \sum_k d(k) \Phi(2t-k) \tag{C.4.2}$$

Beide Gleichungen stellen die entscheidende Verbindung zwischen Wavelets und Filtern dar. Die Eigenschaften der Filterkoeffizienten *c(k)* und *d(k)* werden durch die 2- Skalen Gleichungen von Skalierungsfunktion bzw. Waveletfunktion bestimmt, sie beinhalten jeweils die Argumente *t* und *2t.* Wir beginnen mit dem Raum $V_1 = V_0 \oplus W_0$:

$$\begin{aligned} \sum_k a_{1,k} \Phi_{1,k}(t) &= \sum_k a_{0,k} \Phi_{0,k}(t) + \sum_k b_{0,k} \Psi_{0,k}(t) \\ &= \sum_k a_{0,k} \Phi(t-k) + \sum_k b_{0,k} \Psi(t-k) \end{aligned} \tag{C.4.3}$$

Die Translationen ergeben sich unter Berücksichtigung der 2-Skalengleichungen und der Substitution *l=2k+n*:

$$\Phi(t-k) = \sum_n \sqrt{2}\ c(n)\ \Phi(2(t-k)-n) = \sum_l \sqrt{2}\ c(l-2k) \Phi(2t-l)$$

bzw. (C.4.4)

$$\Psi(t-k) = \sum_n \sqrt{2}\, d(n)\ \Phi(2(t-k)-n) = \sum_l \sqrt{2}\, d(l-2k) \Phi(2t-l)$$

Unter Berücksichtigung von $\Phi(2t-l) = \Phi_{1,l}(t)$ wird nun (C.4.4) jeweils mit dem linken Term von (C.4.1) multipliziert und anschließend über *t* integriert.

$$\int \sum_l c(l-2k) \Phi_{1,l}(t) \sum_l a_{1,l} \Phi_{1,l}(t)\, dt$$

(C.4.5)

$$\int \sum_l d(l-2k) \Phi_{1,l}(t) \sum_l a_{1,l} \Phi_{1,l}(t)\, dt$$

Eine Vereinfachung von (C.4.5) gelingt durch Ausnutzen der Orthogonalitätsbeziehung (C.3.4) und führt direkt auf die Projektionskoeffizienten.

$$a_{0,k} = \sum_l c(l-2k)\, a_{1,l} \qquad b_{0,k} = \sum_l d(l-2k)\, a_{1,l} \qquad \text{(C.4.6)}$$

Nun wird auch die Realisierung der diskreten Wavelet-Transformation als eine rekursive Filterbank mit nachfolgender Unterabtastung um den Faktor zwei deutlich (kritisch abtastende Filterbank).

$$a_{0,k} = \sum_l c[-(2k-l)]\, a_{1,l} \qquad b_{0,k} = \sum_l d[-(2k-l)]\, a_{1,l} \qquad \text{(C.4.7)}$$

Die Koeffizienten *c(k)* und *d(k)* können als Impulsantwort eines diskreten FIR- Tiefpass $H_0(z)$ bzw. FIR- Hochpassfilter $H_1(z)$ aufgefasst werden. Abbildung C.4.1 zeigt einen elementaren Filterblock, mit dem das Signal $a_{j+1,k}$ in Skalierungs- und Wavelet- Koeffizienten der nächst niedrigeren Auflösung zerlegt wird.

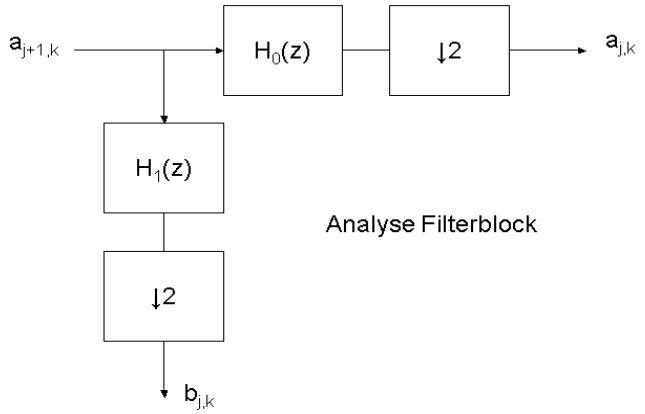

Abbildung C.2: ***Der Analyse Filterblock zerlegt das Signal der Stufe j+1 mittels Hoch- und Tiefpass in Signale der Auflösungstufe j. Da die Filter jeweils die Hälfte der Bandbreite des Signals aus Stufe j+1 aufweisen, kann anschließend eine Unterabtastung um den Faktor vorgenommen werden. Beide Signale der Stufe j decken dann wieder den gesamten Nyquistbereich $0 \le \omega T < \pi$ ab.***

Dieses Schema lässt sich weiterentwickeln, indem der Analyse- Filterblock rekursiv auf die Koeffizienten $a_{j,k}$ angewendet wird. Dabei beginnt der Algorithmus mit einem Signal aus dem Raum V_{J+1} und zerlegt dieses Signal in die orthogonalen Unterräume $V_0 + \sum_J W_j$. Die Ellipsen in Abbildung C.3 verdeutlichen die unterschiedliche lokale Ausbreitung der Wavelets in den einzelnen Skalierungsstufen.

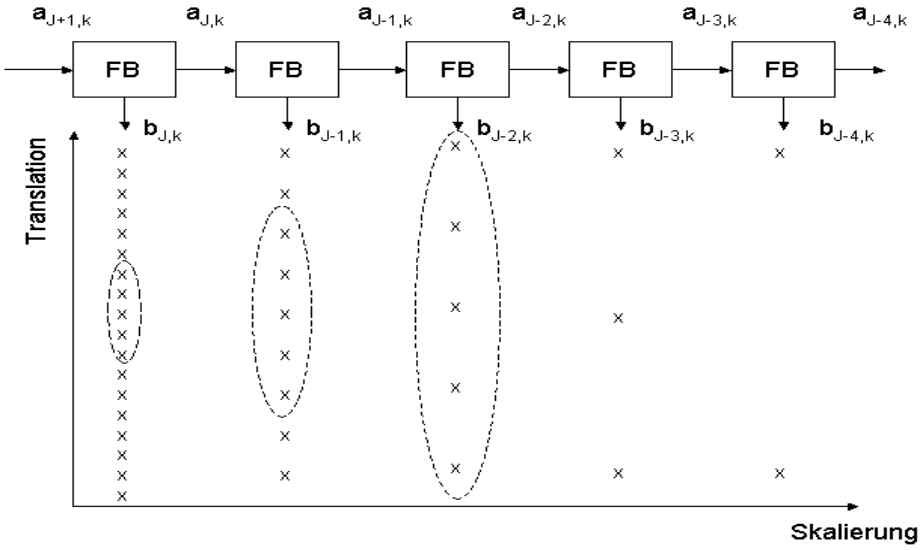

Abbildung C.3: ***Der Zerlegungsalgorithmus entspricht einer rekursiven Anwendung des Basisfilterblocks auf die Skalierungskoeffizienten der jeweiligen Auflösungsstufe.***

C.5 Realisierung der DWT, Berechnung der Filterkoeffizienten

Im Abschnitt C.4 wurde bereits darauf hingewiesen, dass zwischen der diskreten Wavelet-Transformation und Filterbänken mit kritischer Abtastung eine enge Verbindung besteht. Durch die baumförmige Struktur der DWT reduziert sich deren Realisierung auf den Entwurf einer zweikanaligen Analyse/Synthesebank. In diesem Abschnitt werden daher zunächst die Bedingungen für perfekte Rekonstruktion für eine Zweikanalfilterbank herausgestellt. Daraus lassen sich die Beziehungen für alle Übertragungsfunktionen der verwendeten Filter herleiten, eine besondere Bedeutung kommt dabei dem Entwurf des Analyse- Tiefpassfilters zu.

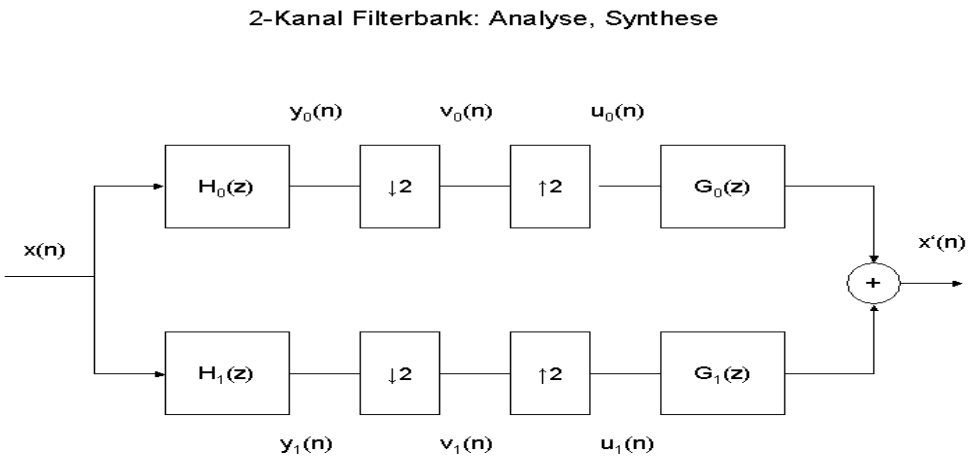

Abbildung C.4: ***Bei der praktischen Realisierung einer Zweikanalfilterbank können auf der Analyseseite keine idealen Filter verwendet werden. Die Frequenzantworten der Filter überlappen sich und es treten Amplituden- bzw. Phasenverzerrungen auf. Die Synthesefilter haben die Aufgabe diese Fehler zu eliminieren.***

Zunächst werden die auftretenden Teilsignale im Frequenzbereich mathematisch beschrieben. Unter Berücksichtigung der Definition für die Kreisfrequenz $\omega = 2\pi f$ und der Abtastfrequenz $f_s = \frac{1}{T}$ können die Teilsignale in den z- Bereich überführt werden. Man beachte, dass mit der normierten Darstellung $\Omega = \omega T$ der Nyquistbereich auf das Intervall $0 \le \Omega < \pi$ abgebildet wird.

Zur Vereinfachung können die Filterblöcke in Abbildung C.3 vorerst weggelassen werden. Die Unterabtastung um den Faktor 2 ist durch die Beziehung

$$y(n) = x(2n) \tag{C.5.1}$$

gegeben, wir beginnen allerdings mit der Beschreibung des Signals u(n):

$$u(n) = \begin{cases} x(n) & , n \text{ gerade} \\ 0 & , n \text{ ungerade} \end{cases} \tag{C.5.2}$$

Mit der Beziehung $e^{-j\pi n} = (-1)^n$ kann man das Signal auch im Frequenzbereich formulieren:

$$U(\omega T) = \frac{1}{2}\sum x(n) e^{-jn\omega T} + \frac{1}{2}\sum x(n) e^{-jn(\omega T + \pi)}$$

oder (C.5.3)

$$U(\omega T) = \frac{1}{2}\left[X(\omega T) + X(\omega T + \pi)\right]$$

Berücksichtigt man (C.5.3) und die Skalierungseigenschaft der Fourier-Transformation:

$$x(2n) \quad \overset{F}{\longleftrightarrow} \quad \frac{1}{2} X\left(\frac{\omega T}{2}\right) \tag{C.5.4}$$

erhält man für die Fourier-Transformierte des unterabgetasteten Signals *v(n)* den folgenden Ausdruck:

$$V(\omega T) = U\left(\frac{\omega T}{2}\right) = \frac{1}{2}\left[X\left(\frac{\omega T}{2}\right) + X\left(\frac{\omega T}{2} + \pi\right)\right] \tag{C.5.5}$$

Unterabtastung kann als Stauchung des Zeitsignals um den Faktor 2 betrachtet werden, dabei verdoppelt sich die Bandbreite des gestauchten Signals. Bei Abtastratenreduktion um den Faktor 2 können bereits an dieser Stelle die Bedingungen für die Übertragungs-funktionen des Tief- bzw. Hochpassfilters angegeben werden:

$$Y_0(\omega T) = 0 \text{ , } für \ \frac{\pi}{2} \le \omega T < \pi \quad \text{ und } Y_1(\omega T) = 0 \text{ , } für \ 0 \le \omega T < \frac{\pi}{2} \tag{C.5.6}$$

Die oben gefundenen Beziehungen werden nun mit $z = e^{j\omega T}$ in den z- Bereich überführt:

$$V(z) = \frac{1}{2}\left[X(z^{\frac{1}{2}}) + X(-z^{\frac{1}{2}}) \right]$$

(C.5.6)

$$U(z) = \frac{1}{2}\left[X(z) + X(-z)\right]$$

Zwischen den z- Transformierten gilt offensichtlich der Zusammenhang $U(z) = V(z^2)$ bzw. $V(z) = U(z^{\frac{1}{2}})$, wobei $z^2 = e^{j2\omega T}$ einer Frequenzverdopplung und $z^{\frac{1}{2}} = e^{j\frac{\omega T}{2}}$ einer Frequenzhalbierung entspricht. Das negative Vorzeichen vor z entspricht bei spektraler und normierter Darstellung einer Verschiebung um π.

$$X(-z) \xrightarrow{F} X(\omega T + \pi) \qquad \text{(C.5.7)}$$

Die in (C.5.6) genannten Forderungen stellen allerdings sehr harte Bedingungen dar, welche für die jeweiligen FIR- Filter eine sehr hohe Anzahl von Koeffizienten zur Folge hätte und darüber hinaus große Signallaufzeiten verursachen würden. Es stellt sich also die Frage, ob die in Abbildung C.3 angegebenen Filter auch bei kurzen Impulsantworten eine perfekte Rekonstruktion ermöglichen. Berücksichtigt man für die Übertragungsfunktionen der Filter eine Verzögerung um *l*-Takte nach folgender Gleichung,

$$x'(n) = x(n-l) \qquad \text{(C.5.8)}$$

so können schnell weichere Bedingungen für perfekte Rekonstruktion angegeben werden.

$$V_0(z) = \frac{1}{2}\left[H_0(z^{\frac{1}{2}})\, X(z^{\frac{1}{2}}) + H_0(-z^{\frac{1}{2}})\; X(-z^{\frac{1}{2}}) \right] \qquad \text{(C.5.9)}$$

$$V_1(z) = \frac{1}{2}\left[H_1(z^{\frac{1}{2}})\, X(z^{\frac{1}{2}}) + H_1(-z^{\frac{1}{2}})\; X(-z^{\frac{1}{2}}) \right] \qquad \text{(C.5.10)}$$

Nach Aufwärtstastung und Filterung mit den Synthesefiltern gelten für die Ausgänge folgende Beziehungen:

$$X'_0(z) = \frac{1}{2} G_0(z)\left[H_0(z)\, X(z) + H_0(-z)\; X(-z)\right] \qquad \text{(C.5.11)}$$

$$X'_1(z) = \frac{1}{2} G_1(z)\left[H_1(z)\, X(z) + H_1(-z)\; X(-z)\right] \qquad \text{(C.5.12)}$$

Die Addition der Teilsignale führt letztlich auf eine ganzheitliche Beschreibung der Filterbank.

$$\begin{aligned} X'(z) &= \frac{1}{2}\left[G_0(z)H_0(z) + G_1(z)H_1(z)\right] X(z) \\ &+ \frac{1}{2}\left[G_0(z)H_0(-z) + G_1(z)H_1(-z)\right] X(-z) \end{aligned} \tag{C.5.13}$$

Der erste Term in Klammern ist für Signalverzerrungen verantwortlich, der zweite hingegen für die Aliaskomponenten.

Hieraus lassen sich zwei Bedingungen für perfekte Rekonstruktion ableiten:

- Verzerrungsfreiheit $\left[G_0(z)H_0(z) + G_1(z)H_1(z)\right] = 2z^{-l}$ (C.5.14)
- Aliasfreiheit $\left[G_0(z)H_0(-z) + G_1(z)H_1(-z)\right] = 0$ (C.5.15)

Die zweite Bedingung wird eingehalten, wenn folgende Beziehungen gelten

$$G_0(z) = H_1(-z);\ G_1(z) = -H_0(-z) \tag{C.5.16}$$

Zur Einhaltung der Verzerrungsfreiheit werden die nachfolgenden Definitionen

$$P_0(z) = G_0(z)H_0(z) \text{ und } P_1(z) = G_1(z)H_1(z) \tag{C.5.17}$$

eingeführt. Hieraus folgt dann unmittelbar die Definitionsgleichung eines Halbbandfilters:

$$P_0(z) - P_0(-z) = 2z^{-l} \tag{C.5.18}$$

Bis auf den zentralen geraden Koeffizienten bei z^{-l} sind alle geraden Koeffizienten Null, die ungeraden Koeffizienten addieren sich zu Null. D.h. für den Entwurf einer Zweikanal-filterbank muss ein Filter $P_0(z)$ gefunden werden, welche die Bedingung (C.5.18) erfüllt. Die Übertragungsfunktionen der Analyse- bzw. Synthesefilter erhält man durch spektrale Faktorisierung. Neben dieser allgemeinsten Entwurfsvorschrift kann außerdem noch das Verfahren der ***Q**uadrature **M**irror **F**ilter* hervorgehoben werden. Bei diesem Verfahren ergibt sich der Betrag des Hochpassfilters $|H_1(\omega T)|$ aus dem Spiegelbild des Betrages des Tiefpassfilters $|H_o(\omega T)|$ bezüglich der halben Nyquistfrequenz $\pi/2$.

$$H_1(z) = H_0(-z) \xrightarrow{z^{-1}} (-1)^n h_0(n) = h_1(n) \tag{C.5.19}$$

Literaturverzeichnis

[Allen-94] Allen, J.B.: *How Do Humans Process and Recognize Speech?*, IEEE Transaction On Speech And Audio Processing, Vol. 2, N0. 4, October 1994.

[Allen-96] Allen, J.B.: *Harvey Fletcher's Role In The Creation Of Communication Acoustics,* J. Acoust. Soc. Am. 99, April 1996.

[Ali-00] Ali, A.M.A.; Spiegel, J.; Mueller, P.: *Auditory-Based Speech Processing Based On The Average Localized Synchrony Detection,* IEEE 2000.

[Andringa-02] Andringa,T.C.: *Continuity Preserving Signal Processing*, Dissertationsschrift 2002, Rijksuniversiteit Groningen.

[Barlow-61] Barlow, H.B.: *Possible Principles Underlying The Transformation Of Sensory Messages,* Sensory Communication, MIT Press, Cambridge MA 1961.

[Bäni-02] Bäni,W.: *Wavelets, Eine Einführung für Ingenieure*, Oldenbourg Verlag 2002.

[Bilmes-98] Bilmes, J. A.: *Maximum Mutual Information Based Reduction Strategies For Cross-Correlation Based Joint Distribution Modelling*, ICASSP 98, April 1998.

[Bourlard-96] Bourlard, H. Dupont, S.: *A New ASR Approach Based On Independent Processing And Recombination Of Partial Frequency Bands,* Int. Conf. On Spoken Language Processing 1996.

[Brayda-04] Brayda, L.; Rigazio, L.; Boman, R.; Junqua, J.C.; *Sensitivity Analysis Of Noise Robust Methods*, ICASSP 2004.

[Bregman-90] Bregman, A.S.: *Auditory Scene Analysis, The Perceptual Organization Of Sound*, Cambridge, MA: MIT Press, 1990.

[Brown-97] Brown, R.G.; Hwang,P.Y.C.: *Introduction To Random Signals and Applied Kalman Filtering,* John Wiley & Sons, 1997.

[Burke-97] Burke, B.: *Wavelets. Die Mathematik der kleinen Wellen,* Birkhäuser Verlag 1997.

[Chi-03] Chi, T.; Ru, P.; Shamma, S.A.: *Multiresolution Spectrotemporal Analysis Of Complex Sound,* Speech Communication, 2003.

[Claes-91] Claes, T.; Compernolle, D. V.: *SNR-Normalisation For Robust Speech Recognition,* ICASSP 1991.

[Cover-91] Cover, T.M.; Thomas, J.A.: *Elements of Information Theory*, John Wiley & Sons, 1991.

[Crouse-98] Crouse, M.S.; Nowak R.D.; R.G. Baraniuk: *Wavelet-Based Statistical Signal Processing Using Hidden Markov Models*, IEEE Transactions On Signal Processing, Vol. 46, NO. 4, April 1998.

[Dau-97a] Dau, T. Kollmeier, B.; Kohlrausch, A.: *Modeling Auditory Processing Of Amplitude Modulation I: Modulation Detection And Masking With Narrowband Carriers,* J. Acoust. Soc. Am. 102, 1997.

[Dau-97b] Dau, T. Kollmeier, B.; Kohlrausch, A.: *Modeling Auditory Processing Of Amplitude Modulation II: Spectral And Temporal Integration In Modulation Detection,* J. Acoust. Soc. Am. 102, 1997.

[Dau-99] Dau, T.: *Modell der effektiven Signalverarbeitung im Gehör,* EINBLICKE Nr. 29, Carl von Ossietzky Universität Oldenburg, April 1999.

[Dean-05] Dean, Th.: *A computational Model of the Cerebral Cortex, 2005.*

[Dimitriadis-02] Dimitriadis, D.; Maragos, P.; Potamianos, A.: *Modulation Features For Speech Recognition,* ICASSP Vol.1, 2002.

[Dimitriadis-05a] Dimitriadis, D.; Maragos, P.; Potamianos, A.: *Auditory Teager Energy Cepstrum Coefficients for Robust Coefficients For Robust Speech Recognition,* Proc. European Conf. on Speech Communication and Technology, Interspeech 2005, Lisbon Portugal, September 2005.

[Dimitriadis-05b] Dimitriadis, D.; Maragos, P.; Potamianos, A.: *Robust AM-FM Features Coefficients For Speech Recognition,* IEEE Signal Processing Letters, Vol.12. No. 9, September 2005.

[Doblinger-95] Doblinger, G.: *Computationally Efficient Speech Enhancement By Spectral Minima Tracking in Subbands,* EuroSpeech'95, Madrid, September 1995.

[Duda-01] Duda, R.O.; Hart P.E.; Stork D.G.: *Pattern Classification, Second Edition*, John Wiley & Sons, Inc. 2001.

[Dusan-05] Dusan, S.; Rabiner, L.R.: *On Integrating Insights from Human Speech Perception into Automatic Speech Recognition,* Interspeech 2005, Lisbon Sept 2005.

[Elhilali-06] Elhilali,M.; Shamma,S.: *A Biologically Inspired Approach To The Cocktail Party Problem,* ICASSP 2006.

[Eska-97] Eska, G.: *Schall und Klang, Wie und was wir hören,* Birkhäuser Verlag 1997.

[Fant-60] *Fant, G.: Acoustic Theory of Speech Production*, Gravenhage, The Netherlands: Mounton and Co., 1960.

[Flanagan-72] Flanagan, J.: *Speech Analysis, Synthesis and Perception*, NY, Springer-Verlag, 1972.

[Fink-03] Fink, G.A.: *Mustererkennung mit Markov-Modellen, Theorie Praxis Anwendungsgebiete*, B. G. Teubner Verlag 2003.

[Fletcher-53] Fletcher, H.: *Speech and Hearing in Communication,* New York: Krieger 1953.

[Fontaine-95] Fontaine, V.; Leich, H.; Ris, C.: *Speech Analysis Based On Malvar Wavelet Transform*, ICASSP 1995.

[Fu-96] Fu, M.; Tan, B.T.; Dermody P.: *The Use Of Wavelet Transforms In Phoneme Recognition, ICSLP 1996.*

[George-05] George, D.; Hawkins, J.: *Invariant Pattern Recognition using Baysian Inference On Hierarchical Sequences,* RNI Technical Report 2005.

[George-06] George, D.; Hawkins, J.: *A Hierchical Bayesin Model Of Invariant Pattern Recognition In The Visual Cortex,* International Joint Conference on Neural Networks 2006.

[Ghitza-94] Ghitza,G.: *Auditory Models and Human Performance In Tasks Related To Speech Coding And Speech Recognition*, IEEE Transaction on Speech and Audio Processing vol. 2, No. 1, 1994.

[Gowdy-00] Gowdy, J.N.; Tufekci, Z.: *Mel Scaled Discrete Wavelet Coefficients For Speech Recognition,* ICASSP 2000.

[Green-91] Green, P. G.; Kuhl, P.K.; Meltzoff, A.N.; Stevens, E.B.: *Integrating Speech Information Across Talkers, Genders and Sensory Modality: Female Faces and Male Voices In The McGurk Effect,* Perception and Psychophysics 50 (6), 524-536, 1991.

[Hansen-95] Hansen J.H.L.; Nandkumar, S.: *Robust Estimation Of Speech In Noisy Backgrounds Based On Aspects Of The Auditory Process,* J. Acoust. Soc. Am., Vol. 97, No. 7, June 1995.

[Hasegawa-97] Jing, Z.; Hasegawa-Johnson, M.: *Auditory Modeling Inspired Methods Of Feature Extraction For Robust Automatic Speech Recognition,* ICASSP 1997.

[Hawkins-04] Hawkins, J.; Blakeslee, S.: *On Intelligence, Times Books, Henry Holt and Company, New York, NY 10011, Sept. 2004.*

[Hermansky-94] Hermansky, H.; Morgan, N.: *RASTA Processing Of Speech,* Speech and Audio Processing, IEEE Transactions on Volume 2, Issue 4, Oct 1994 Page(s): 578 – 589.

[Hermansky-95] Hermansky, H.; Wan, E. A.; Avendano, C.: *Speech Enhancement Based On Temporal Processing*, ICASSP 1995.

[Hermansky-99] Hermansky, H. Sharma, S.: *Temporal Patterns (TRAPS in ASR Of Noisy Speech*, ICASSP 1999.

[Hilberg-97a] Hilberg, W.: *Theorie der Hierarchischen Textkomprimierung. Informationstheoretische Analyse einer deterministischen Sprachmaschine,* Teil 1, Frequenz 51 1997.

[Hilberg-97b] Hilberg, W.: *Theorie der Hierarchischen Textkomprimierung. Informationstheoretische Analyse einer deterministischen Sprachmaschine,* Teil 2, Frequenz 51 1997.

[Hirsch-95] Hirsch H.G.; Ehrlicher C.: *Noise Estimation Techniques For Robust Speech Recognition,* ICASSP 1995.

[Hirsch-00] Hirsch, H.G.; Pearce D.: *The Aurora Experimental Framework For The Performance Evaluation Of Speech Recognition Systems Under Noisy Conditions,* ISCA ASR 2000, Paris France, September 18-20, 2000.

[Hütt-01] Hütt, M.T.: *Datenanalyse in der Biologie*, Springer Verlag 2001.

[Jabloun-95] Jabloun, F.; Cetin, A.E.; Yardimici, Y.: *Subband Analysis For Speech Recognition In The Presence Of Car Noise,* ICASSP 1995.

[Jabloun-99] Jabloun, F.; Cetin, A.E.; Erzin E.: *Teager Energy Feature Parameters For Speech Recognition in Car Noise,* IEEE Signal Processing Letters, Vol. 6, N0. 10, October 1999.

[Junqua-00] Junqua,J.C.: *Robust Speech Recognition in Embedded Systems and PC Applications,* Kluwer Academic Publishers 2000.

[Kaiser-90] Kaiser, J.F.: *On A Simple Algorithm To Calculate The Energy Of A Signal,* ICASSP 1990.

[Kajita-95] Kajita, S.; Itakura, F.: *Robust Speech Feature Extraction Using SBCOR Analysis*, ICASSP Detroit, USA 1995.

[Kajita-96] Kajita, S.; Itakura, F.: *Subband-Crosscorrelation Analysis for Robust Speech Recognition*, ICSLP Philadelphia, USA 1996.

[Kajita-98] Kajita, S.; Itakura, F.: *Spectral Weighting Of SBCOR for Noise Robust Speech Recognition* , ICASSP Seattle, USA 1998.

[Kim-92] Kim, C.W.; Ansari R.; Cetin, A.E.: *A Class Of Linear-Phase Regular Biorthogonal Wavelets,* ICASSP Vol.4 1992.

[Kim-99] Kim, D.S.; Lee, S.Y.; Kil, R.M.: *Auditory Processing Of Speech Signals For Robust Speech Recognition In Real-World Noisy Environment*, IEEE Transaction on Speech and Audio Processing vol. 7, No. 1, January 1999.

[Kryze-99] Kryze, D.; Rigazio, L.; Applebaum T.; Junqua, J.C.: *A New Noise-Robust Subband Front-End And Its Comparison To PLP*, IEEE ASRU Workshop, Keystone, Colorado USA, 1999.

[Lippmann-97] Lippmann, R. P.: *Speech Recognition by Machines and Humans,* Speech Communication 22 (1997) 1-15, 1997.

[Maragos-93] Maragos, P.; Kaiser J.F.; Quatieri, T.F.: *On Amplitude and Frequency Demodulation Using Energy Operators,* IEEE Transactions On Signal Processing, Vol. 41, No. 4, April 1993.

[Marr-82] Marr, D.: *Vision: A computational investigation into the human representation and processesing of visual information,* San Francisco : W.H. Freeman, 1982.

[Martens-90] Martens, J.P.: Immerseel, L.V.: *An Auditory Model based On The Analysis Of Envelope Patterns,* ICASSP 1990.

[Martin-93] *Martin,R.: An Efficient Algorithm To Estimate The Instantaneous SNR Of Speech Signals,* Eurospeech 1993.

[Martin-94] *Martin, R.: Spectral Subtraction Based On Minimum Statistics,* Proc. IEEE Signal Processing Sept. 1994.

[Martin-01] Martin, R.: *Noise Power Spectral Density Estimation Based On Optimal Smoothing and Minimum Statistics*, IEEE Transaction on Speech and Audio Processing, Vol. 9. No. 5, July 2001.

[Mesgarani-05] Mesgarani, N.; Shamma, S.: *Speech Enhancement Based On The Filtering The Spectrotemporal Modulations,*ICASSP Vol.1 2005.

[Mirghafori-98] Mirghafori, N.; Morgan, N.: *Transmission And Transitions: A Study Of Two Common Assumptions In Multi-Band ASR,* ICASSP Vol.2 1998.

[Mirghafori-99] Mirghafori, N.; Morgan, N.: *Sooner Or Later: Exploring Asynchrony In Multi-Band Speech Recognition,* Eurospeech 1999.

[Mountcastle-78] Mountcastle, V.:, *An Organizing Principle for Cerebral Function: The Unit Model and the Distributed System*, The Mindful Brain (Gerald M. Edelman and Vernon B. Mountcastle, eds.) Cambridge, MA: MIT Press 1978.

[Mumford-03] Lee,T.S.; Mumford. D.: *Hierarchical Bayesian inference in the visual cortex,* Journal of the Optical Society of America, Vol. 20, No.7, July 2003.

[Nilsson-00] *Nilsson*, M.; Andersen, S. V.; Kleijn, W. B.: *On The Mutual Information Between Frequency Bands In Speech,* ICASSP 2000.

[Paliwal-98] *Paliwal K.K.: Spectral Subband Centroid Features For Speech Recognition, ICASSP 1998.*

[Papoulis-02] Papoulis, A.; Pillai, S.: *Probability, Random Variables, and Stochastic Processes*, McGraw-Hill Higher Education; Auflage: 4th ed., Januar 2002.

[Perdigao-97] Perdigao, F.S.; de Sa, L.V.: *Properties Of Auditory Model Representations*, Proc. EuroSpeech'97, Vol. 5, Sept. 1997.

[Perdigao-98] Perdigao, F.S.; de Sa, L.V.: *Auditory Models As Front Ends For Speech Recognition Representations*, NATO ASI on Computational Hearing, Il Ciocco, Italy 1998.

[Perdigao-99] Perdigao, F.S.; de Sa, L.V.: *A Noise Suppression Technique Using An Auditory Model,* EuroSpeech'99, Sept. 1999.

[Peters-98] Peters, S.D.; Stubley, P.; Valin, J.M.: *On The Limits Of Speech Recognition In Noise*, ICASSP 1998.

[Pollock-99] Pollock, D.S.G.: *Time-Series Analysis, Signal Processing And Dynamics*, Academic Press Ltd., 1999.

[Potamianos-96] Potamianos, A.; Maragos, P.: *Speech Formant Frequency and Bandwidth Tracking Using Multiband Energy Modulation*, Journal of Acoustical Society of America 99,June 1996.

[Ramach.-99] Ramachandran, V.S.: *Phantoms in the Brain: Human Nature and the Architecture of the Mind*, Fourth Estate 1999.

[*Rieke-99]* Rieke, F.; Warland, D.; Van Stevenick, R.; Bialek, W.: *Spikes, Exploring The Neural Code*, MIT Press, 1999.

[Rizolatti-96] Rizzolatti, G.: Premotor cortex and the recognition of motor actions, Cognitive Brain Research 1996.

[Rizolatti-06] Rizzolatti, G.; Fogassi, L.; Gallese, V.: *Mirrors in the Mind,* Scientific American Band 295, Nr. 5, November 2006.

[Römer-00] Römer, R.; Burchard, B.: *A Single Chip Phoneme Based HMM Speech Recognition System for Consumer Applications*, IEEE Conference Proceedings, ICCE 2000.

[Römer-02] Römer, R.; Koloska U.; Hirschfeld, U.: *Optimierung der Erkennleistung von HMM basierten Spracherkennern*, ESSV Dresden, September 2002.

[Römer-06] Römer, R.; Brückner, R.: *Vergleichende Untersuchungen zur Zuverlässigkeit von Pitch-kohärenten Merkmalen bei verschiedenen Störgeräuschen unter Verwendung der Aurora-2 Datenbasis*, ESSV Freiberg, August 2006.

[Römer-07] Römer, R.: *Vergleichende Untersuchungen zur Erkenngenauigkeit Pitch-kohärenter Merkmale bei verschiedenen Störgeräuschen unter Verwendung der Aurora-2 Datenbasis*, ESSV Cottbus, September 2007.

[Ruske-94] Ruske, G.: *Automatische Spracherkennung, Methoden der Klassifikation und Merkmalsextraktion,* Oldenbourg Verlag 1994.

[Seltzer-03] Seltzer, M.; Droppo J.; Acero, A.: *A Harmonic-Model-Based Front End For Robust Speech Recognition,* Eurospeech 2003.

[Seneff-84] Seneff, S.: *Pitch And Spectral Estimation Of Speech Based On Auditory Synchrony Model*, Proc. IEEE International, ICASSP 1984.

[Shamma-85] Shamma, S.A.: *Speech Processing in the Auditory System I: The Representation Of Speech Sounds In The Response Of The Auditory Nerve,* J. Acoust. Soc. Am. 78, November 1985.

[Shamma-85] Shamma, S.A.: *Speech Processing in the Auditory System II: Lateral Inhibition And The Central Processing of Speech Evoked Activity In The Auditory Nerve,* J. Acoust. Soc. Am. 78, November 1985.

[Shamma-03] Shamma, S.: *Encoding Sound Timbre In The Auditory System,* IETE Journal of Research, Vol. 49, April 2003.

[Stahl-00] Stahl, V.; Fischer A.; Bippus R.: *Quantile based Noise Estimation For Spectral Subtraction And Wiener Filtering,* ICASSP 2000.

[Strang-97] Strang, G.; Nguyen, T.: *Wavelets and Filterbanks*, Wellesley-Cambridge Press 1997.

[Tchorz-99] Tchorz,J.; Kollmeier, B.: *A Model Of Auditory Perception As Front End For Automatic Speech Recognition,* JASA 106 (4), October 1999.

[Terhardt-98] Terhardt, E.: *Akustische Kommunikation, Grundlagen mit Hörbeispielen*, Springer Verlag, 1998.

[Thomlinson-97] Thomlinson,M.J.; Rusell,M.L.: *Modelling Asynchrony In Speech Using Elementary Single-Signal Decomposition,* ICASSP Vol.2 1997.

[Vereecken-95] Vereecken, H.; Martens, J.P.: *Recognition Of Noisy Speech Using An Auditory Model,* EuroSpeech'95, 1995.

[Vereecken-96] Vereecken, H.; Martens, J.P.: *Noise Supression And Loudness Normalization In An Auditory Model-Based Acoustic Frontend,* ICSLP '96, pp.566-569, 1996.

[Viikki-1997] Viikki, O.; Laurila, K.: *Noise Robust HMM-Based Speech Recognition Using Segmental Feature Vector Normalization,* ESCA- NATO Workshop on Robust Speech Recognition For Unknown Communication Channels, Pont-a-Mousson, France 1997.

[Viikki-1998] Viikki, O.; Bye, D.; Laurila, K.: *A Recursive Feature Vector Normalization Approach For Robust Speech Recognition In Noise,* IEEE 1998.

[Vondra-07] Vondra, M.; Vich, R.: *Adaptive Comb Filtering In Speech Enhancement By Spectral Subtraction,* ESSV Cottbus, September 2007.

[Voicebox] http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html.

[Wang-93] Wang, K.; Shamma, S.A.; Byrne, W.J.: *Noise Robustness In The Auditory Representation Of Speech Signals,* ICASSP 1993.

[Wang-94a] Wang, K.; Shamma, S.A.: *Self-Normalization And Noise Robustness In Early Auditory Representations*, IEEE Transaction On Speech And Audio Processing, Vol. 2, N0. 3, 1994.

[Wang-94b] Wang, K. Shamma, S.A.: *Zero-Crossings and Noise Suppression in Auditory Wavelet Transformations,* Technical Research Report TR 92-94, University of Maryland 1994.

[Weinrichter-91] Weinrichter, H.; Hlawatsch, F.: *Stochastische Grundlagen nachrichtentechnischer Signale*, Springer Verlag, 1991.

[Yang-92] Yang, X.; Wang, K.; Shamma, S. A.: *Auditory Representations Of Acoustic Signals,* IEEE Transactions On Information Theory, Vol. 38, March 1992.