Unterdrückung hörbarer Störgeräusche in Echtzeitsystemen

Diplomarbeit durchgeführt am
Institut für Elektronische Musik und Akustik

Durchführung: Franz Zotter

Betreuung: Robert Höldrich, O. Univ-Prof. Dr. techn. DI. Mag.
Markus Noisternig, DI .

Graz, Oktober 2004

Kurzfassung

Um die Störsignalunterdrückung einer Freisprecheinrichtung möglichst verzerrungs- und verzögerungsfrei zu gestalten, soll in dieser Diplomarbeit eine auditive Gammatone Filterbank verwendet werden. Die Filterbank lässt sich im Zeitbereich implementieren, ist somit nicht an Blocksignalverarbeitung gebunden. Dadurch entfallende Signalbuffer helfen, die Latenzzeit zu verringern. Der wesentliche Vorteil der auditiven Signalanalyse ist, dass sie ohne Zusatzmaßnahmen die Tonhöhenwahrnehmung und die Simultanverdeckung des menschlichen Ohres modelliert.
Zusätzlich wird durch einfache Mittel die zeitliche Vor- und Nachverdeckung des Gehörs nachgebildet. Die auf die Gammatone-Filterbank Analyse basierende Rauschunterdrückung und Störsignalschätzung bewirkt in Folge nur mehr die Wiederherstellung der Mithörschwelle des ungestörten Signals und beschränkt sich somit auf die Unterdrückung hörbarer Störungen. Auf diese Weise werden unnötige Signalverzerrungen gering gehalten und musical noise verringert.

Abstract

To avoid distortion and processing delay in a hands free communication system, this work uses an auditory Gammatone filter bank instead of a fourier transform (FFT). The filter bank can be fully implemented in the time domain and thus doesn't need block processing. Additional singal buffering can be omitted completely, which helps in reducing processing delay. The major advantage of auditory signal analysis is that it contains pitch perception and simultaneous masking of the human ear inherently. Temporal pre and post masking models can be achieved with simple additional effort. As a consequence, the Gammatone filter bank based noise reduction and noise estimation only reconstructs the masking threshold of the undisturbed signal, so its effect is reduced to the suppression of audible noise. In this way the influence on the signal is kept as low as possible and musical noise is further reduced.

Inhaltsverzeichnis

1 Einleitung

In einem am Institut für Elektronische Musik und Akustik laufenden Projekt wird versucht ein *Multi-Sensor-System für Sprachaufbesserung* zu implementieren, das in einer *Freisprecheinrichtung* eingesetzt werden und folgende wesentlichen Funktionen beinhalten soll:

1. Erkennung der Stimmaktivität und der Einfallsrichtung mithilfe der Signale zweier Mikrofone (VAD, DOA):
   a) Liegt ein aktives Sprachsignal vor? - Detektion über ein entsprechend angepasstes Energiemaß. (voice activity detection - VAD)
   b) Kommt das Sprachsignal aus der richtigen Richtung? - Detektion der Verzögerung eines aufgezeichneten Signals zwischen zwei örtlich getrennten Mirkofonen. (direction of arrival - DOA)
2. Räumliche Filterung durch nachgeführte Mikrofonzeilen-Richtcharakteristik (*array-processing, beamforming*):
   a) Sprecheraktivität: Fokussierung der Mikrofonzeilen-Richtwirkung auf den *aktiven Sprecher* im erlaubten Aktionsradius zur Aufbesserung des Sprachsignals
   b) Sprecherinaktivität, bzw. Interferenzsprecher/quelle aktiv: Fokussierung der Mikrofonzeilen-Richtwirkung auf akustische Störungen, um eine gezieltere Unterdrückung zu ermöglichen. (*Störgeräuschquellen, interferierende Sprecher* außerhalb des erlaubten Aktionsradius)
3. Störgeräuschunterdrückung:
   a) Psychoakustisch verbesserte Unterdrückung von bekannten Störungen, welche durch nachgeführte Mikrofonrichtwirkung extrahiert werden können. Dabei sollen Eigenschaften des Gehörs genützt werden, um statt der Subtraktion der ermittelten Störsignale eine verzerrungsarme Unterdrückung in einer psychoakustischen Domäne durchzuführen.
   b) Unterdrückung von Störungen, die durch nachgeführte Richtwirkungen nicht extrahiert werden konnten durch Verfahren zur spektralen Subtraktion. Wird diese spektrale Subtraktion in einer psychoakustischen Domäne durchgeführt, können dabei entstehende Nutzsignalverzerrungen gering gehalten werden.
4. Akustische Aufbesserung der Eigensprachverständlichkeit zur Vermeidung des *Lombard-*

Reflexes[1] mit aktiver Rückkopplung:

a) Dabei ist vor allem wichtig, dass der Rückkopplungspfad Lautsprecher-Mikrofon (*LEM, loudspeaker-enclosure-microphone*) mit einer wirksamen Kompensation akustischer Echos (*AEC, acoustic echo cancellation*) unterbunden wird.

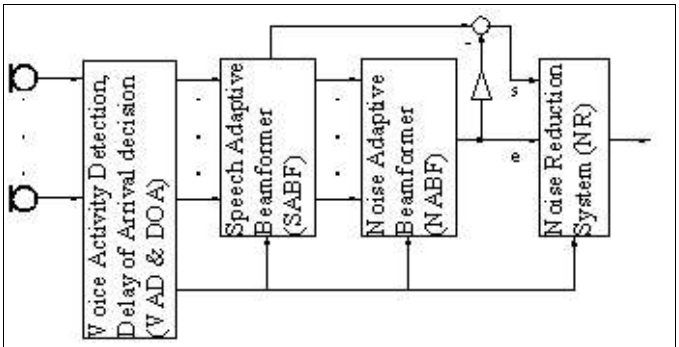

Abbildung 1: Systemüberblick

In dieser Arbeit soll speziell das Teilgebiet der Unterdrückung *wahrnehmbarer Störgeräusche* untersucht werden, wobei die Effizienz der Implementierung ein wesentliches Kriterium darstellt.

- Implementierung im Zeitbereich: Das Gesamtsystem soll ohne FFT (schnelle Fourier-Transformation) realisiert werden. Die benötigte Spektralanalyse soll mithilfe auditiver Filterung und Pegeldetektion im Zeitbereich durchgeführt werden.
- Latenzarme Implementierung: Die Verarbeitungsstufe soll in Echtzeit funktionieren und möglichst wenig Verzögerung des Nutzsignals verursachen.
- Ressourceneffizienz: Zur Störsignalunterdrückung nötige Berechnungen sollen mit sehr wenig Rechenleistung auskommen. Dabei muss sowohl ein günstiger *trade-off* zwischen Qualität und Rechenersparnis, als auch zwischen Rechenersparnis und Latenzzeit (Downsampling!) erzielt werden.
- Psychoakustisches Modell: Wenn nur wahrnehmbare Störungen entfernt werden, wird der Eingriff ins Nutzsignal – und somit die Nutzsignalverzerrung – möglichst gering gehalten:
  - Simultanverdeckung des Gehörs: Leise Störungen beinhaltende Spektralkomponenten, die aufgrund eines lauten Nutzsignals in der unmittelbaren spektralen Umgebung nicht als Störung wahrgenommen werden können, sollen nicht unterdrückt werden. Dadurch wird einer Verzerrung der wahrnehmbaren spektralen Gestalt des Nutzsignals vorgebeugt.

1 Der Lombard-Reflex beschreibt das Phänomen unbewusster Anpassung der Sprechlautstärke, um umgebende Störgeräusche zu kompensieren [1].

- Zeitliche Verdeckung des Gehörs: Störungsereignisse, die aufgrund iherer zeitlichen Nähe zu lauten Nutzsignalereignissen nicht hörbar (zeitlich verdeckt) sind, sollen ebenfalls möglichst erhalten bleiben, um die zeitliche Struktur des Nutzsignals nicht zu beschädigen.
- Auditory Scene Analysis: Gruppierungsvorgänge im menschlichen Gehör, die akustisch vermischte Schallereignisse aufgrund gemeinsamer Modulationsstrukturen von Frequenz und Amplitude erkennen und trennen können, sollen auch als Anhaltspunkt dienen, um eine zusätzliche Verbesserung des Signal-Stör-Verhältnisses zu verwirklichen. Es wird sich möglicherweise nur eine Einschränkung der Amplitudenmodulationsfrequenzen effizient umsetzen lassen.

Vorhandene Arbeiten das Gesamtsystem:

Die Arbeiten von Markus Noisternig, Cornelia Falch [2] am Institut für Elektronische Musik und Akustik, sowie die Patentschrift [3] von S. K. Hui dienen hier als Vorlage, in welcher die beschriebene Teilfunktion optimiert werden soll. Wichtige IEEE-Veröffentlichungen in diesem Themengebiet sind die Arbeiten von X. Zhang und J. H. L. Hansen [4] (*beamforming*), R. Balan und J. Rosca [5] (*beamforming & noise suppression*).

Verwendete Arbeiten zur Psychoakustik:

Pflüger [6][7], Zwicker [8][9], Terhardt [10], Lin *et al* [11][12][13], Lyon [14][15], Kubin *et al* [16], Slaney [17][18], PEAQ [19], Baumgarte [20], Irino *et al* [21][22][23], Thiemann [24], etc.

Verwendete Arbeiten zur (Computational) Auditory Scene Analysis:

Ellis [25], Bregman [26], Scheirer [27], Unoki *et al.* [28], Slaney [18], Wang *et al*. [29].

Verwendete Arbeiten zur Störgeräuschunterdrückung:

Ephraim und Malah [30][31], Cohen [32][33][34][35], Cappé [36], Höldrich *et al* [37], Vary/Heute/Hess [38], Tsoukalas *et al* [39], etc.

2 Beschreibung des menschlichen Gehörs

2.1 Außen-Mittelohr

Dieser Abschnitt soll als kurze Beschreibung des Übertragungsverhaltens von Außenohr und Innenohr dienen.

2.1.1 Außenohr

Die Übertragung des Schallfelds durch das Außenohr setzt sich aus folgenden Komponenten zusammen. Zum Einen treten Färbungen des Luftschalls am menschlichen Körper, Kopf, Torso und der Ohrmuschel in Abhängigkeit zur Schalleinfallsrichtung auf. Zum Anderen tritt auch eine wesentliche Klangfärbung durch die Resonanzfrequenzen des Ohrkanals, der von der Ohrmuschel zum Trommelfell führt, auf. Der Kopf und das Außenohr führen dabei zu einer Wellenanpassung zwischen Schallfeld und Gehörgang. Die Färbung durch die Einfallsrichtung zur Vereinfachung oft nur in zwei Fälle unterteilt (vgl. Zwicker [8], Terhardt [10]):

5. Schalleinfall aus Blickrichtung (0° - Freifeldkurve)
6. Schalleinfall aus unbekannter/diffuser Richtung (Diffusfeldkurve)

2.1.2 Mittelohr

Das schräg im Ohrkanal liegende Trommelfell leitet den Luftschall weiter ins luftgefüllte Mittelohr, wo die Trommelfellschwingung durch die Übersetzung der Gehörknöchelchen impedanztransformiert an die Flüssigkeiten des Innenohrs angekoppelt werden. Die Gehörknöchelchen besitzen aufgrund ihrer vorhandenen Masse und Steifigkeit eine Bandpasscharakteristik. Zwei an die Knöchelchen angreifende Muskel können die Mittenfrequenz dieses Bandpass leicht zu höheren Frequenzen hin verschieben um das Innenohr zu schützen, benötigen dazu aber eine Reaktionszeit von etwa 100 ms. Das Öffnen der Eustachischen Röhre zum Druckausgleich zwischen Mittel- und Außenohr geschieht nur beim Schlucken und Gähnen [10][9].

2.1.3 Form der Außen-Mittelohr-Übertragungsfunktion

Die Form einer Außen-Mittelohr-Übertragungsfunktion kann aus den Kurven gleicher Lautstärke (KGL) qualitativ erkannt werden, welche als Freifeldkurven (0° Schalleinfallsrichtung) gemessen wurden. Die KGL zeigen die Empfindlichkeit des Ohres bei Anregung mit harmonischen Schallschwingungen bestimmter Frequenz (Sinus) und beinhalten somit unter anderem die

Auswirkungen der Außen-Mittelohr-Übertragungsfunktion. Zum Beispiel ist die Resonanz des Ohrkanals als erhöhte Empfindlichkeit, also als Einbuchtung in den Kurven gleicher Lautstärke bei etwa 3-4 kHz zu sehen [10]. Die Frage bleibt allerdings:

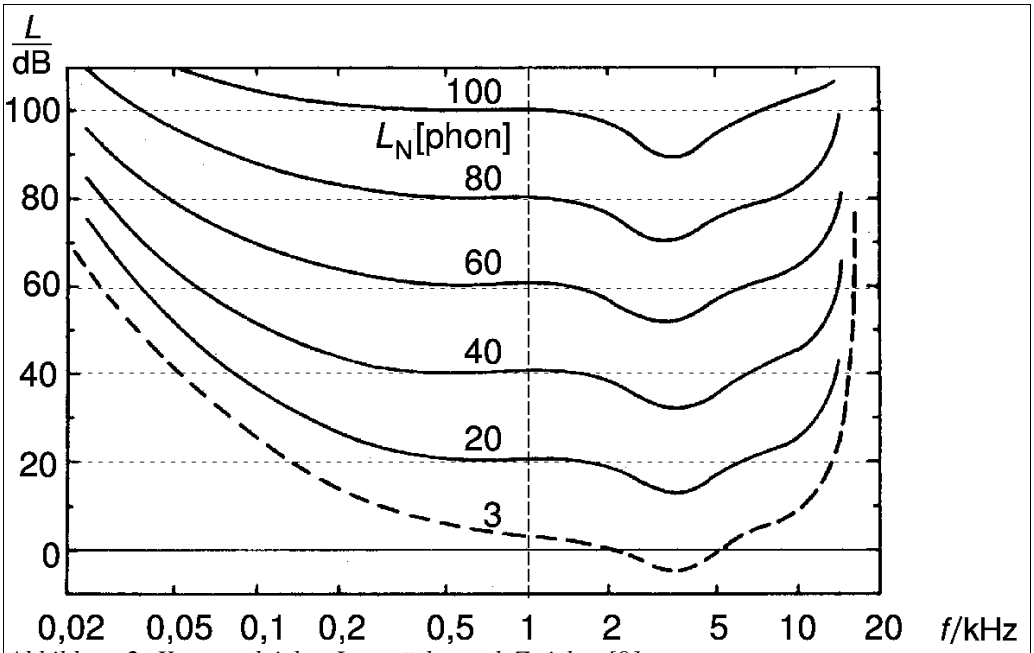

Abbildung 2: Kurven gleicher Lautstärke nach Zwicker [9]

Welche Phonkurven besitzen in jeweils welchem Abschnitt die Form eines inversen Außen-Mittelohrfilters?

Folgende Punkte sind der Literatur zu entnehmen (Zwicker [9] und Pflüger [6]):

- Die Phonkurven verlaufen mit (Ausnahme großer Lautstärken) über 1 kHz nahezu parallel.
- Unter 1 kHz sind die Phonkurven nicht parallel, bei leisen Pegeln sind die Flanken steiler.
- Die Ruhehörschwelle ist wahrscheinlich in der größten Genauigkeit bekannt, weshalb ihr Verlauf über 1 kHz als inverse Außen-Mittelohr Übertragungsfunktion verwendet werden kann.
- (Der Anstieg der Empfindlichkeit in ISO-Kurven im Bereich unter 1 kHz wurde in den modifizierten Phonkurven, welche aktuellere Messungen der KGL in gemittelter Form enthalten, widerlegt [6].)
- Die bioelektrische Aktivität der Haarzellen im Innenohr und der Blutkreislauf erzeugen ein niederpegeliges tieffrequentes Rauschen, das bei sehr geringen Lautstärkeverhältnissen zu höheren Frequenzen hin verdeckend wirkt [10]. Ab etwa 100 Phon ist die verdeckende Wirkung dieses Rauschens zu vernachlässigen [9]. Deshalb kann für die Beschreibung einer inversen

Außen-Mittelohr Übertragungsfunktion die 100 Phonkurve bei Frequenzen unter 1 kHz verwendet werden.

Die gesuchte inverse Filterform kann somit gefunden werden, indem der Kurvenzug der (modifizierten) 100 Phonkurve für Frequenzen unterhalb 1 kHz mit jenem der Ruhehörschwelle im Bereich höherer Frequenzen zusammengefügt wird.

2.2 Innenohr

In diesem Abschnitt wird die Funktionsweise des menschlichen Innenohres beschrieben. Des weiteren wird skizziert, wie diese Funktion in einem technischen Modell nachgebildet werden kann.

2.2.1 Funktion des Innenohres

Schallwellen werden durch das Außenohr aufgenommen und treffen auf dem Trommelfell auf, über welche sie ins Mittelohr gelangen. Vom Mittelohr werden die Schallwellen auf mechanischem Wege über die Gehörknöchelchen ins Innenohr geleitet (ovales Fenster). Das Innenohr besteht aus einem schneckenförmigen Gehäuse (Kochlea), dessen gewundener innerer Hohlraum in Längsrichtung in 3 mit 2 unterschiedlichen Flüssigkeiten befüllte Kammern geteilt ist (Scala vestibuli, Scala media, Scala tympani). Auf den die Flüssigkeiten trennenden Membranen (Reissnersche Membran, Basilarmembran) bildet sich bei Anregung durch das Mittelohr (am ovalen Fenster) eine Wanderwelle aus (Wanderwellentheorie von Békésy, vgl. [8][10]). Diese Wanderwelle bildet ein Amplitudenmaximum aus, das bei tiefen Anregungsfrequenzen tiefer in der Gehörschnecke liegt als bei hohen Anregungsfrequenzen. Auf einer der Membranen (Basilarmembran) angesiedelte passive Nervenzellen (innere Haarzellen), die mit einer Deckmembran (Tektorialmembran) über Härchen verbunden sind, werden durch eine Scherbewegung der Wanderwellenausbreitung zur Aussendung von elektrischen Impulsen angeregt. Haarzellen, welche in der Nähe eines Amplitudenmaximums angesiedelt sind, werden am stärksten angeregt und ermöglichen bei der Auswertung des Anregungsmusters im Gehirn die Tonhöhenwahrnehmung (Frequenz-Orts-Transformation des inneren Ohres). Es ist zudem nicht auszuschließen, dass die Haarzellen ihrerseits eine Frequenzselektivität besitzen [10].

Neben passiven inneren Haarzellen befinden sich auch aktive äußere Haarzellen auf der Basilarmembran im Innenohr. Über diese Nervenzellen kann eine aktive Regelung der Empfindlichkeit durchgeführt werden, welche einerseits die Empfindlichkeit bei leisen Signalen erhöhen und andererseits die Frequenzselektivität der Basilarmembran durch Entdämpfung steigern

[10]. Bei gesunden Ohren ist die Antwort dieser Regelschleife auf einen kurzen Impulsschall am Ohrkanal messbar (otoakustische Emissionen, [8]). Die bei dieser Rückkopplung entstehenden Nichtlinearitäten führen auch zu wahrnehmbaren Kombinationstönen [10].

In psychoakustischen Experimenten wurden gleichzeitig auftretende Verdeckungseffekte zwischen Tönen unterschiedlicher Frequenz und Amplitude gemessen (Simultanmaskierung), die sich physiologisch auch über die Wanderwellentheorie erklären und modellieren lassen [8]. Treten gemeinsam 2 Töne auf, kann bei bestimmten Frequenz- und Laustärkeverhältnissen zwischen den Tönen nur der lautere Ton wahrgenommen werden. Liegt der maskierte Ton bei höherer Frequenz als sein Maskierer, treten die Verdeckungseffekte stärker in Erscheinung (*tuning curves* [8]).

Der Amplitudenverlauf der Wanderwelle bei Anregung durch einen Einzelton steigt vom Beginn der Basilarmembran ausgehend kontinuierlich bis zum Punkt des Amplitudennmaximums an, wonach er relativ schnell wieder verebbt. Tritt eine Anregung durch einen weiteren Ton auf, dessen Amplitudenmaximum so klein ist, das es bei der Überlagerung der entstehenden Wanderwelle mit jener des anderen Tones nicht mehr detektierbar ist, kann dieser verdeckte Ton auch nichtmehr wahrgenommen werden. Wie bereits erwähnt, treten Wanderwellenmaxima hoher Anregungsfrequenzen am Beginn der Basilarmembran und jene tiefer Frequenzen weiter am Ende der Basilarmembran auf. Weil die Wanderwellenamplitude einen flachen Anstieg vom Beginn der Basilarmembran bis zum Amplitudenmaximum besitzt und dann rasch wieder verebbt, ist die Verdeckung hoher Frequenzen stärker. Tiefe Frequenzen werden im Gegensatz dazu nur wenig verdeckt.

2.2.2 Frequenzgruppeneigenschaft des Gehörs

Das menschliche Gehör besitzt die Eigenschaft Intensitäten akustischer Reize innerhalb von schmalbandigen Frequenzbändern zu gruppieren und zur Bildung der Lautstärkeempfindung aufzusummieren. Am besten ist diese Eigenschaft anhand der Experimente an der Ruhehörschwelle zu erklären [8].

Wird ein schmalbandiges Rauschen an der Ruhehörschwelle des Gehörs nach jeder Verdoppelung seiner Bandbreite so eingestellt, dass es wieder an der Ruhehörschwelle zu liegen kommt und gerade noch wahrgenommen werden kann, so wird Folgendes beobachtet [8]:

- Bei jeder Verdoppelung der Bandbreite muss der Pegel des Rauschens zunächst um 3 dB gesenkt werden,

- überschreitet die Bandbreite des Rauschens die Frequenzgruppenbreite, kann der Pegel nicht weiter gesenkt werden, da jene Schallintensitäten außerhalb der Frequenzgruppenbreite nichtmehr zur Bildung der Lautstärkeempfindung aufsummiert werden. Die so gefundene Frequenzgruppenbreite $\Delta f_{g,\,CB}$ wird kritische Bandbreite genannt, Zwicker [8].

Das selbe Experiment kann mit eng benachbarten Tönen durchgeführt werden, deren Anzahl stetig verdoppelt wird, die Ergebnisse sind ähnlich, die kritische Bandbreite ist [10]:

$$\frac{\Delta f_{g,CB}}{[Hz]} = 25 + 75\left[1 + 1{,}4\left(\frac{f}{1000}\right)^2\right]^{0{,}69}. \tag{1}$$

Eine Näherung dieses Zusammenhangs ist gegeben durch [9]:

$$\frac{\Delta f_{g,CB}}{[Hz]} \simeq max\,(0{,}2 \cdot f\,,\,100\,Hz). \tag{2}$$

Aus den etwas aufwendigeren Messungen auditiver Filterformen mit der *notched-noise method*[2] (Rosen und Baker, vgl [6]) ist eine weitere Quantifizierung der Frequenzgruppenbreite entstanden, die im gesamten Frequenzbereich geringere Bandbreiten als die kritischen Bandbreiten ergibt. Dabei wurden die energieäquivalenten rechteckigen Bandbreiten[3] (ERB, *equivalent rectangular bandwidth*) der vermessenen auditiven Filter als Anhaltspunkt verwendet (Glasberg+Moore, vgl Pflüger [6] und Terhardt [10]):

$$\frac{\Delta f_{g,ERB}}{[Hz]} = 24{,}7\,(1 + \frac{4{,}3 \cdot f}{1000}). \tag{3}$$

An die Frequenzgruppeneigenschaft des Gehörs sind zwei Transformationen der linearen Frequenzachse in psychoakustische Frequenzdomänen angelehnt:

1. *Tonheit*: Mit einer Transformation $Tfm_{\mathrm{Bark}}\{\cdot\}$ kann die lineare Frequenz f der Einheit *Hz* in die *Tonheit* z_{Bark} der Einheit *Bark* umgerechnet werden. In der Skala der *Tonheit*, oder auch *Bark*-Skala, entspricht einer kritischen Bandbreite genau 1 *Bark.* Die Umrechnung nach Zwicker lautet (vgl [40][10]):

2 Bei der *notched-noise method* wird die Hörschwelle eines Sinustons vermessen, dessen Frequenz mittig in einer schmalbandigen Einkerbung gleichmäßig verdeckenden Breitbandrauschens (Maskierer) liegt. Wird die ermittelte Pegelschwelle in Abhängigkeit zur Einkerbungsbreite aufgetragen erhält man die Form symmetrischer auditiver Filter. Werden von der Einkerbungsmitte abweichende Frequenzen zugelassen, können allgemeinere unsymmetrische auditive Filterformen gefunden werden, indem jene mit der kleinsten Hörschwelle ausgewählt werden.

3 Wird die Fläche unter dem Frequenzgang eines Filters auf das Leistungsdichtespektrum eines rechteckigen Schmalbandrauschens aufgeteilt, dessen Höhe dem Filterpeak entspricht, besitzt dieses Rauschen die energieäquivalente rechteckige Rauschbandbreite (ERB).

$$\frac{z}{[Bark]} = Tfm_{Bark}\left\{\frac{f}{[Hz]}\right\} = 13\arctan\left(0{,}76\frac{f}{1000}\right) + 3{,}5\arctan\left[\left(\frac{f}{75000}\right)^2\right]. \tag{4}$$

Die Umrechnung kann auch mit der einfachen Formel von Traunmüller (vgl. [6][10]) erfolgen:

$$\begin{aligned} \frac{z}{[Bark]} &= Tfm_{Bark}\left\{\frac{f}{[Hz]}\right\} \simeq \frac{26{,}81\frac{f}{1000}}{1{,}96+\frac{f}{1000}} - 0{,}53, \\ \frac{f}{[kHz]} &= Tfm^{-1}_{Bark}\left\{\frac{z}{[Bark]}\right\} \simeq 1{,}96\frac{z+0{,}53}{26{,}28-z}. \end{aligned} \tag{5}$$

2. Die *ERB-Skala* ist qualitativ ähnlich und entspricht einer logarithmischen Frequenz-Skala mit additivem Offset [10], weist aber bei tieferen Frequenzen wesentlich schmälere Frequenzbänder auf als die *Tonheit*:

$$\begin{aligned} \frac{z}{[ERB]} &= Tfm_{ERB}\left\{\frac{f}{[Hz]}\right\} = 21{,}4\cdot\log_{10}\left(1+4{,}37\cdot f\right), \\ \frac{f}{[kHz]} &= Tfm^{-1}_{ERB}\left\{\frac{z}{[ERB]}\right\} = \frac{10^{\frac{z}{21{,}4}}-1}{4{,}37}. \end{aligned} \tag{6}$$

2.2.3 Technische Modelle des Innenohres

Die Frequenzanalyse des menschlichen Gehörs wäre zugänglich, wenn man die Auslenkung der Wanderwelle an ausreichend vielen ortsdisktreten Punkten entlang der Basilarmembran abtasten bzw. die elektrischen Signale einzelner Nervenzellen an diesen Stellen anzapfen könnte. In der Praxis kann statt dessen ein Modell der Wanderwelle verwendet werden, deren Schwingung für einen beliebigen Punkt der Basillarmembran berechnet werden kann [8][10]. Die Übertragungsfunktionen des Schalls hin zu einer diskreten Anzahl von Punkten kann verwendet werden, um Implementierungen einer Kochlea mittels Bänken elektronischer oder digitaler Filter zu beschreiben. Ein leicht verständliches Modell, das nur die passiv angeregte Basilarmembran betrachtet, ist beispielsweise in der Arbeit „Cochlea Mechanics Demystified“ von Richard F. Lyon und Carver Mead [14] zu finden.

Im Wesentlichen reicht eine endliche Zahl solcher kochleären Übertragungsfunktionen aus um die Spektralanalyse des Gehörs zu modellieren, wie in den Modellen von Zwicker oder Terhardt [8][10]. Der Simultanmaskierung entsprechend, enthalten Signale solcher stark überlappenden Filterbänder auch jeweils Komponenten tieferer Frequenz, die maskierend wirken können.

Eine offene Frage dabei bleibt jedoch die Aufteilung der Frequenzbänder entlang des hörbaren Spektrums. Da psychoakustische Messungen, die den Verdeckungseffekt ausnützen, um auf kochleäre Filterformen zu kommen (Rosen & Baker, vgl. [6]), mit der Wanderwellentheorie vereinbar sind, ist es hier sinnvoll die Filterbandbreiten durch Mittel der Psychoakustik zu definieren [8][10].

Meist bedient man sich bei der Festlegung der Mittenfrequenzen der Bark-Skala [8] oder der ERB-Skala (Slaney [17]), welche beide an die Frequenzgruppeneigenschaften des Gehörs angelehnt sind. (0,6 Bark wird unter anderem als ideale Bandbreite für auditive Filter angesehen; für den Hörbereich von 100-8000 Hz wären bei 0,6 Bark-Breite etwa 32 Bänder nötig)

Es kann sogar gezeigt werden, dass zwischen der *Bark*-Skala und der Position auf der Basilarmembran ungefähr ein lineares Verhältnis besteht [8].

Ein kompaktes, leicht parametrisierbares Modell für kochleäre Filterung ist die Gammatone-Filterbank, von der es zahlreiche effiziente Implementierungen, aber auch nichtlineare Implementierungsformen mit dynamischen Eigenschaften gibt. Wesentliche Arbeiten zum Thema Gammatone-Filterung sind Lyon [15], Slaney [17], Ambikairajah [13], Pflüger [7], Irino und Patterson [21]. Die Gammatone-Filter können mit den Messungen psychoakustischer *tuning curves* und Messungen an der Basilarmembran in Einklang gebracht werden.

2.2.4 Auditive Filterformen

Es gibt zahlreiche Methoden zur Formbestimmung auditiver Filter, welche zur kochleären Frequenzanalyse verwendet werden können. Eine einfache Darstellung gemessener auditiver Filterformen bieten die *roex*(*p,r*)-Funktionen (siehe [7], [6]). Mit den gemessenen Wertepaaren können die Parameter p und r gefunden werden. Die Messungen von Rosen und Baker ergeben Filterformen in Abhängigkeit zur Filtermittenfrequenz f_c und dem Signalpegel P_s :

$$roex(p,r) = (1-r)(1+p\cdot g)e^{-p\cdot g} + r, \quad mit \quad g = \frac{|f - f_c|}{f_c},$$

$$p = \begin{cases} p_l = 39 - 0{,}42 \cdot Ps & , f < f_c \\ p_u = 27{,}1 & , f \geq f_c \end{cases}, \quad r = \begin{cases} r_l = 10^{-4.67 + 0.042 \cdot Ps} & , f < f_c \\ r_u = 0 & , f \geq f_c \end{cases}. \tag{7}$$

Diese Filterformen gelten für den Bereich, in dem die Frequenzgruppen des Gehörs logarithmisch aufgeteilt sind (> 500 Hz). Für tiefere Frequenzen sind diese Filterformen zu schmalbandig, man

kann jedoch durch die Frequenzgruppeneigenschaften des Gehörs die Filterformen in einer Frequenzgruppenskala (*Bark*, *ERB*) linear nach oben oder unten verschieben, um die gehörrichtigen Formen für alle Frequenzen zu bekommen [13].

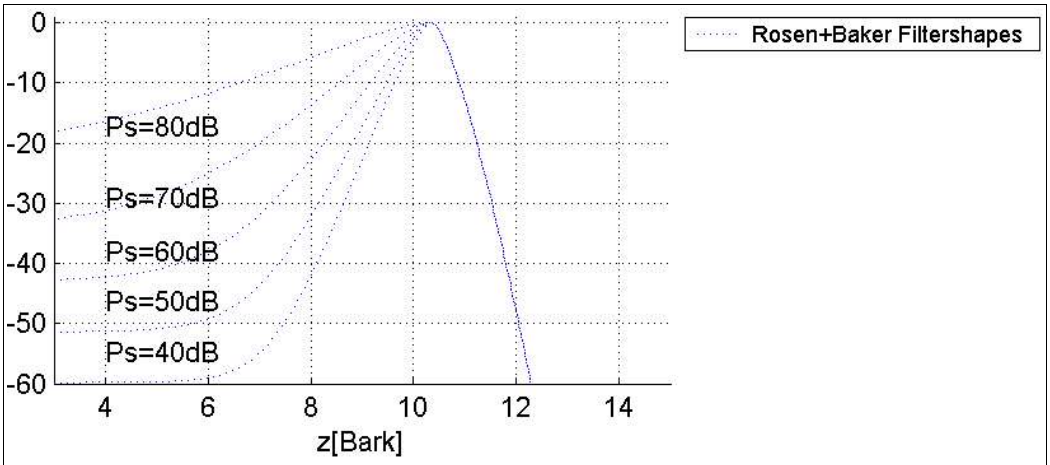

Abbildung 3: Filterformen von Rosen und Baker bei unterschiedlichen Pegeln im Bereich der Sprachlautstärke.

Die oben angegebenen Filter besitzen eine -3 dB Bandbreite von etwa 0.6 Bark und eine -6 dB Bandbreite von etwa 0,9 Bark. Für die Unterteilung des Hörbereiches zwischen 100 Hz und 8 kHz werden ca. 32 Frequenzbänder bei -3 dB, oder 20 Frequenzbänder bei -6 dB Überlappung der Durchllasskurven benötigt. Diese Angaben sind jedoch abhängig vom Lautstärkepegel. Ein lineares Modell würde nur eine Filterkurve für jedes Frequenzband verwenden, ein nichtlineares würde die Pegelabhängigkeit auch nachbilden.

2.3 Aktivität der Nervenzellen, Nachverarbeitung im Gehirn

Die inneren Haarzellen, welche zur Umwandlung von akustischen Schwingungen auf der Basilarmembran des Innenohres in elektrische Impulse fähig sind, feuern bei großer mechanischer Anregung mit einer größeren Rate Impulse ab als bei geringer mechanischer Beanspruchung, wobei die Haarzellen Diodencharakter besitzen, d.h. sie sind speziell auf die Auslenkung in eine Richtung empfindlich [10] (Transduktionsprozess der inneren Harzellen). Die Ladungskapazität einer Nervenzelle ist jedoch nicht unerschöpflich, weshalb sie nach jeder Erregung Zeit zur Regeneration benötigt [10]. Zudem gibt es Effekte, die es sehr lauten Schallereignissen ermöglichen im Nervensystem schneller ausgewertet zu werden als leiseren. Durch das Zusammenwirken dieser

zeitlichen Faktoren, vermutlich auch durch die Aktivität der äußeren Haarzellen, entstehen psychoakustisch messbare Verdeckungseffekte, welche die Maskierung in zeitlichen Abfolgen von Schallereignissen kennzeichnen [8].

Bei der Auswertung von umgebenden Schallfeldern können auch sehr komplexe Vorgänge im Gehirn dafür verantwortlich gemacht werden, dass dicht gemischte Schallquellen getrennt wahrgenommen werden können und auch bei lautem Störgeräuschpegel oft noch Sprache verstanden werden kann. Diese komplexen Vorgänge werden mit den Stichworten „Auditory Scene Analysis“ oder „cocktail-party effect“ bezeichnet.

2.3.1 Vorverdeckung

Die Vorverdeckung des Gehörs beschreibt, wie kurze leise Schallereignisse durch unmittelbar darauf folgende laute Schallereignisse verdeckt, also nicht mehr wahrnehmbar, sind. Die Vorverdeckung ist individuell stark unterschiedlich und von der Art der Schallereignisse abhängig. Die Vorverdeckung ist in etwa bei allen Frequenzen gleich und funktioniert bis maximal 10 ms vor einem Schallereignis, sie kommt vermutlich durch eine begünstigte schnellere Verarbeitung von lauten Schallereignissen im Gehirn zustande [8][10].

2.3.2 Nachverdeckung

Die Nachverdeckung des Gehörs ist von der Art der Schallereignisse abhängig, aber individuell nicht so stark unterschiedlich wie die Vorverdeckung. Sie beschreibt, wie leise Schallereignisse nach lauteren Ereignissen verdeckt werden können und kann bis maximal 200 ms nach einem Schallereignis gemessen werden. Im Gegensatz zur Vorverdeckung ist hier eine Frequenzabhängigkeit der Nachverdeckungsdauer zu erkennen. Demnach ist in den höchsten Frequenzbändern eine nur 4 ms lange Nachverdeckungsdauer zu messen [8][10]. Eine mögliche Begründung der Nachverdeckung kann die begrenzte Kapazität der inneren Haarzellen und die Aktivität der äußeren Haarzellen sein.

2.3.3 Gruppierung – Auditory Scene Analysis (ASA)

Nach Albert Bregman's „Auditory Scene Analysis“ [26] erkennt das menschliche Gehör der Gestalt von akustischen Signalen, um sie trennen und einzuordnen zu können. Zur Charakterisierung solcher Gestalten gibt es mehrere Merkmale:

- spektrale Form und Lautstärke:

- Formanten
- ILD: *interaural level difference* (Richtungswahrnehmung)
- Auswertung der Dynamik und Lautstärke einer Schallquelle

- zeitliche Struktur:
  - Amplitudenmodulation: *common onset/offset*, ...
- zeitliche Feinstruktur: Correlogram, Kreuzkorrelation
  - Frequenzmodulation: *comodulation effects* (gemeinsame Modulation der Periodizität in der Grundschwingung)
  - ITD: *interaural time difference* (Richtungswahrnehmung, Gesetz der 1. Wellenfront)
  - IGD: *interaural group delay* (frequenzselektive Detektion von Laufzeitunterschieden)

Aus der Arbeit von Dan Ellis [25] sind sogenante *wefts* bekannt, welche als akustische Gestalten aus Diagrammen des Schallpegels in einem 3-dimensionalen Raum gewonnen werden können. Dieser Raum entsteht, wenn *Korrelogramme* zu unterschiedlichen Zeitpunkten aneinandergereiht werden.

Korrelogramm (*Correlogram*): Ein Korrelogramm besteht aus den Autokorrelationsfunktionen über die halbwellengleichgerichteten Signale (Transduktion der inneren Haarzellen) in den Frequenzbändern einer kochleären Filterbank. Anhand eines Korrelogramms sind die Periodizitäten der Einzelfrequenzbänder erkennbar. Enthalten benachbarte Frequenzbänder Harmonische unterschiedlicher Grundfrequenzen, passen die Autokorrelationsmuster nicht zusammen. Eine Zuordnung der Frequenzbänder zu den beiden Grundfrequenzen, und damit eine Unterscheidung beider Klänge, ist möglich. Im Gegensatz dazu ergibt das Korrelogramm eines reinen Klanges (eine Grundfrequenz) ein homogenes harmonisches Muster. Ein Gemisch inkohärenter Klänge, die zum Analysezeitpunkt dieselebe Frequenz haben, kann über das Korrelogramm nicht getrennt werden.

Erst durch die Analyse der zeitlichen Abfolge von Korrelogrammen wird es möglich komplexe Klanggemische oder Geräusche anhand von gemeinsamen Frequenz- und Amplitudenmodulationen zu unterscheiden. Dazu genügt die Erkennung von Mustern, sogenannten *wefts*, in der Abfolge von Korrelogrammen, anhand welcher eine Gruppierung und Unterscheidung von Schallquellen getroffen werden kann.

Aufgrund der unterschiedlichen Frequenz- und Amplitudenmodulationen ist es mit dieser aufwendigen Analyse meist möglich, zwei oder mehrere Sprecher voneinander zu unterscheiden. Man nimmt an, dass das menschliche Gehör ebenfalls in der Lage ist solche komplexen Analysen

durchzuführen.

Da der Mensch 2 Ohren besitzt ist eine weitere Verbesserung der Erkennung unterschiedlicher Schallquellen durch ihre Richungsinformation möglich. Richtungsinformationen werden im menschlichen Gehör durch Auswertung von Pegel- und Laufzeitdifferenzen zwischen beiden Ohren, sowie Klangfärbungen, die von der Schalleinfallsrichtung abhängen, extrahiert.

Es existieren bereits technische Modelle dieser Analysemöglichkeiten, die zum Trennen von Schallquellen, zum Unterdrücken von Störungen und zum Erkennen von Tonhöhe eingesetzt werden können. Die computerbasierte Durchführung solcher Analysen wird *computational auditory scene analysis* (CASA) genannt.

Interessante Arbeiten im Themenkreis: Slaney *et al.* beschreiben, wie eine Korrelogrammanalyse wieder resynthetisiert werden kann [18], E. D. Scheirer [27], Wang und Brown [41], Unoki und Agaki [28] beschreiben Verfahren zur Separation von Schallquellen .

3 Digitale Implementierung einer auditiven Signalanalyse

Ein wesentliches Ziel in diesem Abschnitt ist es, eine einfache zeitdiskrete Signalanalyse zu erarbeiten, welche den Eigenschaften der auditiven Signalanalyse im menschlichen Gehör ähnlich ist. Als Grundlage für die digitale Signalverarbeitung wird das Buch von Oppenheim Schafer und Buck [42] empfohlen. In den folgenden Abschnitten werden grundlegende Zusammenhänge über z-Transformation, digitale Filterstrukturen, Pol-Nullstellendiagramme, Frequenz- und Phasengänge als gegeben betrachtet.

3.1 Digitaler Außen-Mittelohr Filter

In diesem Abschnitt wird eine einfache digitale Implementierung von invertierbaren Außen-Mittelohr Übertragungsfiltern beschrieben, die zur Vorverarbeitung von akustischen Signalen verwendet werden können.

Ein vollständiges Modell des Gehörs sollte auch die Übertragungsfunktion des äußeren und mittleren Ohres und die Ruhehörschwelle berücksichtigen, siehe Pflüger [6].

In den Kurven gleicher Lautstärke ist die Wirkung der Außen-Mittelohr-Übertragung zu erkennen, die sich in der Messung der Kurven gleicher Lautstärke als Veränderung der Empfindlichkeit entlang der Frequenzachse bemerkbar macht (freies Schallfeld, Schalleinfall von 0° zum Hörer). Die Außen-Mittelohr Übertragungsfunktionen für das diffuse Schallfeld können durch die Anwendung eines Diffusfeldfilters auf jene Übertragungsfunktionen für das Freifeld bestimmt werden.

Da die Kurven gleicher Lautstärke bei kleinen Signalpegeln Maskierungseigenschaften durch tieffrequentes inneres Rauschen des Ohres (spontane Aktivitäten der Nervenzellen, Blutzirkulation) beeinflusst werden, wird für Frequenzen < 1 kHz die modifizierte inverse 100 Phon-Kurve, und für Frequenzen > 1 kHz die inverse Ruhehörschwelle als Vorgabe zur Bestimmung des Außen-Mittelohr-Filters im Freifeld verwendet (siehe Pflüger [6] und Zwicker [9]).

Gesucht wird nun eine invertierbare Übertragungsfunktion, die als Außen-Mittelohr-Filter eingesetzt werden kann. Dazu werden 3 biquadratiche Filterstufen in Kaskade verwendet (Abbildung 4).

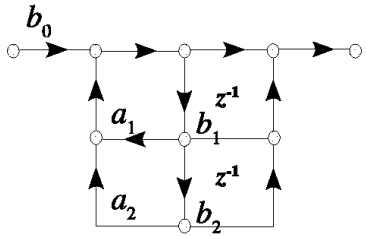

Abbildung 4: Biquadratische digitale Filterstruktur (second order section, SOS)

3.1.1 Hochpasscharakteristik bei tiefen Frequenzen

Die Hochpasswirkung des Ohres, die sich aus dem inneren Rauschen des Ohres und der unteren Hörgrenze ergibt lässt sich durch einen Hochpass mit einer doppelten reellen Nullstelle, die tiefe Frequenzen unterdrücken soll und einer doppelten reellen Polstelle, welche einen Frequenzgang mit *unity gain* bei höheren Frequenzen ermöglichen soll zusammensetzen. Als Optimierungsziel wird die modifizierte inverse 100-Phonkurve verwendet. Die Übertragungsfunktion hat die Gestalt ([6])

$$H_{AMO,fL}(z)=\frac{\left(1-r_{1,fL}z^{-1}\right)^2}{\left(1-r_{2,fL}z^{-1}\right)^2}. \tag{8}$$

Sollte der Nullstellenradius $r_{1,fL}$ den Wert 1 annehmen, könnte dieser für eine stabile Inversion des Filters problemlos auf 0,995 abgeändert werden, ohne dabei wesentliche Signalfärbungen zu erzeugen.

Der Rechenaufwand umfasst 4 Additionen, 4 (2 bei $r_{1,fL} = 1$) Multiplikationen, 2 Speicherstellen und 2 unterschiedliche Koeffizienten.

3.1.2 Tiefpasscharakteristik bei hohen Frequenzen

Die Flanke für hohe Frequenzen wird ähnlich jener für tiefe Frequenzen erzeugt, es sind im Prinzip nur die Vorzeichen der Radien $r_{1,fH}$ und $r_{2,fL}$ unterschiedlich. Der Frequenzgang sollte möglichst gut mit der Kurvenform der inversen Ruhehörschwelle übereinstimmen.

$$H_{AMO,fH}(z)=\frac{\left(1+r_{1,fH}z^{-1}\right)^2}{\left(1+r_{2,fH}z^{-1}\right)^2}. \tag{9}$$

Normalerweise ergeben sich bei der Approximation der gesuchten Tiefpasscharakteristik in diesem Bereich der Außen-Mittelohr Übertragungsfunktion keine kritischen Nullstellenradien in der Nähe

von 1. Der Amplitudengang der Außen-Mittelohr Übertragungsfunktion besitzt hier keinen steilen Verlauf, wie etwa bei tiefen Frequenzen. Daher ist eine Inversion hier unproblematisch. Der Rechenaufwand umfasst 4 Additionen, 4 Multiplikationen, 2 Speicherstellen und 2 unterschiedliche Koeffizienten.

3.1.3 Resonanz bei mittlerer Frequenz

Ein Peakfilter zur Nachbildung der Resonanz des Gehörgangs kann sehr einfach mithilfe eines Allpasses 2. Ordnung zusammengesetzt werden. Dabei wird der Ausgang des Allpass so vom Eingangssignal subtrahiert, dass bei der Phasendrehung um 180° *unity gain* und bei den Phasen 0° und 360° eine vollständige Auslöschung auftritt. Das so erzeugte Bandpasssignal wird wiederum dem Eingangssignal gewichtet hinzuaddiert, damit sich bei der Mittenfrequenz eine lokale Anhebung bzw. Absenkung gegenüber dem umliegenden *unity gain* ausbildet (siehe DAFX [43]).
Die Übertragungsfunktion sollte der inversen Ruhehörschwelle im Bereich von 3-4 kHz angenähert werden:

$$H_{AMO,res}(z)=\frac{\left[\frac{(1+r_{res}^2)(1-g)}{2}+g\right]-2r_{res}\cos(\theta_{res})z^{-1}+\left[\frac{(1+r_{res}^2)(1+g)}{2}-g\right]z^{-2}}{1-2r_{res}\cos(\theta_{res})z^{-1}+r_{res}z^{-2}}. \tag{10}$$

Dabei steuert r_{res} die Bandbreite und θ_{res} die Mittenfrequenz der Resonanz, und mit g kann der Betrag der Überhöhung bzw. der Absenkung im Vergleich zu *unity gain* eingestellt werden.
Die Inversion dieses Filters ist auch unproblematisch, da der Faktor g niemals 0 oder ∞ sein wird.
Der Rechenaufwand beträgt 4 Additionen, 5 Multiplikationen, 2 Speicherstellen und 4 unterschiedliche Koeffizienten.

3.1.4 Die gesamte Außen-Mittelohr-Übertragungsfunktion

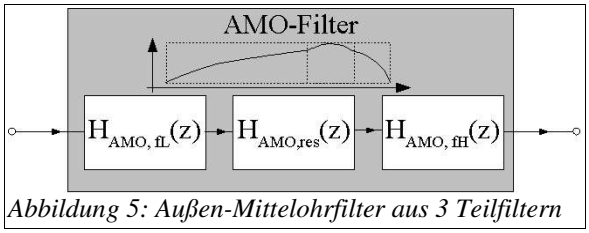

Abbildung 5: Außen-Mittelohrfilter aus 3 Teilfiltern

In Abbildung 5 ist ein Blockschaltbild des vollständigen Außen-Mittelohrfilters zu sehen und in Abbildung 6 der damit erreichte Frequenzgang im Vergleich zu den aus den Kurven gleicher

Lautstärke zusammengesetzten Kurventeilen. Die Tiefpassflanke bei hohen Frequenzen kann bei bandbegrenzten Systemen meist weggelassen werden. Für noch gröbere Näherungen ist es durchaus möglich nur den Teilfilter mit der Hochpassflanke bei tiefen Frequenzen zu verwenden. Das ermittelte Parameterset zu den gefundenen Kurven ist in der untenstehenden Tabelle zu sehen. Durch die Überlappung der einzelnen Filtersegmente ändern sich die Parameter je nach Näherung geringfügig.

	$r_{1,fL}$	$r_{2,fL}$	r_{res}	θ_{res}	g_{res}	$r_{1,fH}$	$r_{2,fH}$
f_s=22,05 kHz	1 bzw. 0,995	0,976	0,66	1,044	2,72	0,51	0,16
	1 bzw. 0,995	0,976	0,66	0,98	2,72	0	0
	1 bzw. 0,995	0,976	0	0	0	0	0
f_s=16 kHz	1 bzw. 0,955	0,97	0,65	1,577	3,72	0,55	0,35
	1 bzw. 0,955	0,97	0,65	1,42	3,36	0	0
	1 bzw. 0,955	0,97	0	0	0	0	0

Insgesamt beträgt der Rechenaufwand eines Außen-Mittelohr Filters mit Hochpass-Flanke und Resonanz etwa 17 Operationen.

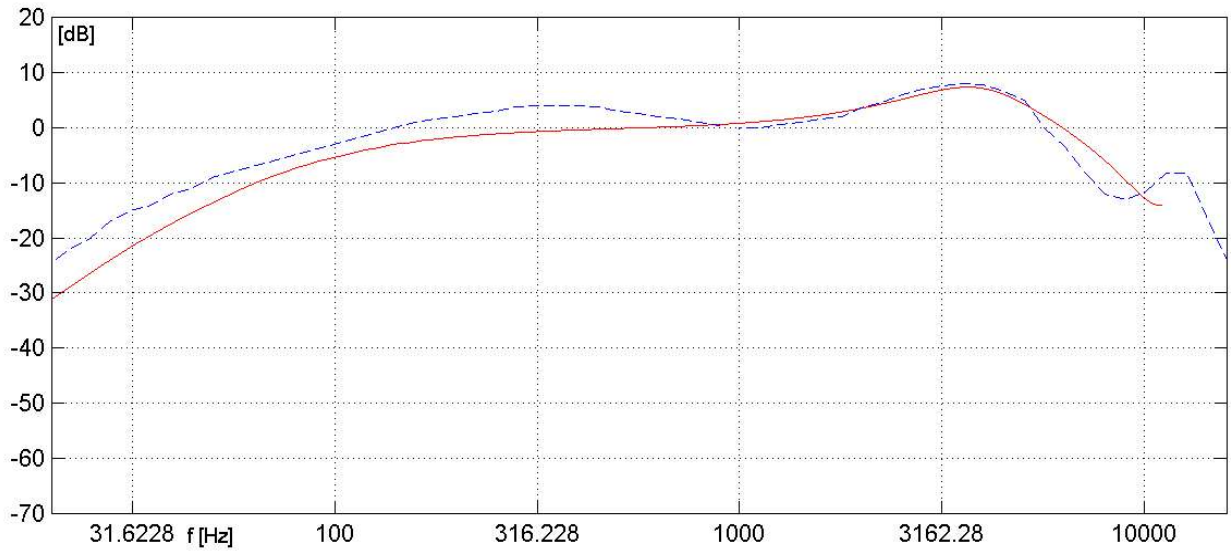

Abbildung 6: Nachbildung des Außen-Mittelohr-Filters mit 3 Teilfiltern.
(gestrichelt: inverse 100 Phon-Kurve ($f < 1$ kHz) mit inverser Ruhehörschwelle ($f > 1$ kHz) kombiniert und normalisiert; durchgezogen: parametrisch eingestellter Außen-Mittelohr Übertragungsfilter)

3.2 Frequenzanalyse: Gammatone-Filter

In diesem Abschnitt soll beschrieben werden, wie eine Filterbank aus parallelen Gammatone-Filtern zur Nachbildung von psychoakustischen Maskierungseffekten und den physiologischen Gegebenheiten des Innenohres verwendet werden kann. Gammatone-Filter lassen sich recht gut mit

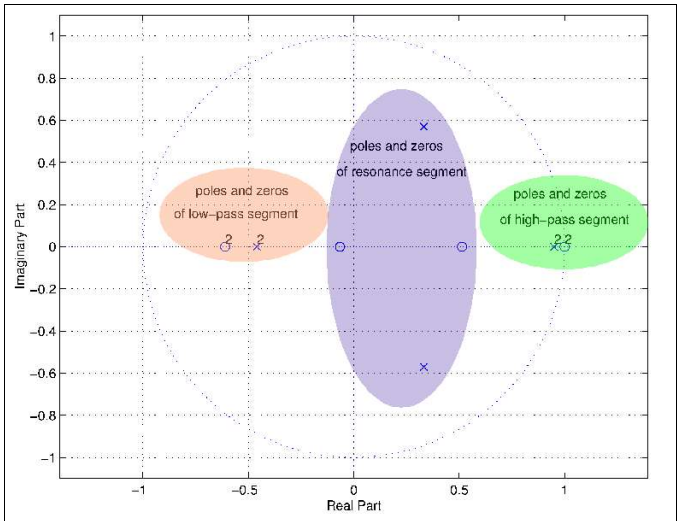

Abbildung 7: Pol- Nullstellendiagramm eines Außen-Mittelohr-Filters im z-Bereich

den Beschreibungen der hydrodynamischen Wanderwelle im Innenohr, den Daten aus psychoakustischen Experimenten, aber auch mit physiologisch gemessenen Impulsantworten in Übereinstimmung bringen und sind damit qualitativ gut geeignet, die Spektralanalyse des menschlichen Gehörs zu modellieren [7] [6]. Die Verwendung von Gammatone-Filtern ist nicht zuletzt deshalb attraktiv, weil dafür effiziente Implementierungen mit niedriger Latenz im Zeitbereich existieren [17]. Vom Gammatone abgeleitet lassen sich der All-Pole Gammatone, One-Zero Gammatone [15], ein „Three-Zero“ Gammatone [13] und der Gammachirp [44] finden, die zum Teil eine bessere Übereinstimmung mit den psychoakustischen *tuning curves* und effiziente Implementierung bieten.

3.2.1 Mittenfrequenzen und Überlappung

Zur Aufteilung des Hörbereichs in Bandpasssignale kann entweder die ERB-Skala (energieäquivalente rechteckige Rauschbandbreite auditiver Filter) oder die Bark-Skala (kritische Bandbreiten, CB) verwendet werden, um die Frequenzgruppeneigenschaft des Gehörs nachzubilden.

Für eine möglichst effiziente Implementierung mit einer groben Frequenzeinteilung zu bekommen ist die *Bark*-Skala nach Traunmüller am geeignetsten, da die so erhaltenen Bandpässe von allen möglichen Skalen mit der geringsten Güte auskommen.

Allgemein kann die Aufteilung des Frequenzbereichs in N Bänder des Index k mit den Mittenfrequenzen $f_c(k)$ und den Bandbreiten B_w zwischen f_{Lo} und f_{Hi} bei gegebener Transformation $Tfm\ \{\cdot\}$ in eine psychoakustische Frequenzskala nach folgender Formel erfolgen:

$$f_c(k) = Tfm^{-1}\left\{ Tfm\left\{f_{Lo}\right\} + \frac{Tfm\left\{f_{Lo}\right\} - Tfm\left\{f_{Hi}\right\}}{N-1} \cdot k \right\} \quad \text{, mit k = 0 , 1, ... , N-1} \tag{11}$$

$$B_w(k) = Tfm^{-1}\left\{f_c(k+0{,}5)\right\} - Tfm^{-1}\left\{f_c(k-0{,}5)\right\}. \tag{12}$$

Mit B_w kann die Bandbreite der einzelnen Frequenzbänder unabhängig von der verwendeten Transformation berechnet werden.

Mit den auditiven Filterformen nach Rosen und Baker (vgl [6]) besitzt eine Filterbank mit -3 dB Dämpfung an den Überlappungsfrequenzen im Bereich zwischen 100 Hz und 8 kHz (1 Bark bis 20 Bark) etwa 33 Kanäle. Bei einer Überlappung mit -6 dB Dämpfung wären 20 Frequenzbänder ausreichend.

3.2.2 Lineare Gammatone-Filter (GF)

Eine Herleitung der gewöhnlichen Gammatone-Filter ist in [17] zu finden. Dazu wird einfach die bekannte Impulsantwort des Gammatones in den Laplace-Bereich transformiert, von wo aus man relativ einfach auf elektronische und digitale Implementierungsformen kommen kann. Die Gammatone-Impulsantwort der Ordnung M wird beschrieben durch die Gleichung:

$$h_{M,GF}(t) = g_{M,GF} \cdot t^{M-1} e^{-bt} \cos(\omega t). \tag{13}$$

In den Laplace-Bereich transformiert – anders als im zeitdiskreten z-Bereich – erhält man eine einheitliche Darstellung für alle Ordnungen M:

$$H_{M,GF}(s) = g_{M,GF} \cdot \frac{(-1)^{M-1}(M-1)!}{2} \frac{(s+b-j\omega)^M + (s+b+j\omega)^M}{\left[(s+b)^2+\omega^2\right]^M}. \tag{14}$$

Der Zähler der Übertragungsfunktion lässt sich relativ leicht in reelle Polynome 2. Grades faktorisieren (*second order sections*), wobei der Nenner bereits in faktorisierter Darstellung vorhanden ist. Um eine zeitdiskrete Repräsentation zu erhalten ist es am einfachsten die Impuls-

Invarianz-Technik auf die *second order sections* der Laplace-Darstellung anzuwenden[4]. Es kann gezeigt werden, dass der zeitdiskretisierte Gammatone ausreichend bandbegrenzt ist, sodass sich bei ausreichendem Abstand der Mittenfrequenz $\omega/(2\pi)$ von der Bandgrenze $f_s/2$ eine Übertragungsfunktion ohne Aliasing-effekte findet [17]. Die Berechnungen der Null- und Polstellen der Gammatone-Funktionen etlicher Ordnungen und die Berechnungen der Bandbreite, Überlappung, etc. sind dem Anhang zu entnehmen.

Die Laplace- oder z-Bereichsbeschreibung einer Gammatone-Funktion besitzt jeweils M konjugiert komplexe Polstellenpaare und M relle Nullstellen. Wird eine Sinus-Schwingung als Trägerschwingung verwendet, reduziert sich die Anzahl der reellen Nullstellen auf (M - 1). Die genaue Form kann dem Anhang entnommen werden. Das Pol-Nullstellendiagramm im digitalen z-Bereich eines Gammatones ist in Abbildung zu sehen. Die Vielfachpolstellen sind dabei entsprechend der Mittenfrequenz und Bandbreite des Gammatone-Filters vorgegeben, die M Nullstellen auf der reellen Achse sind von diesen Vorgaben abhängig.

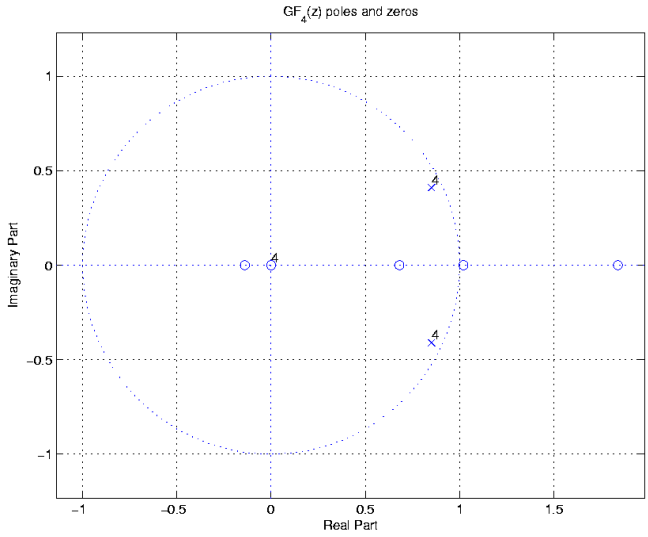

Abbildung 8: Pol-Nullstellen-Diagramm eines 4. Ordnung Gammatone-Filters mit cos-Phase.

In der Literatur wird meist der Gammatone 4. Ordnung verwendet, mit dem Bandbreiteparameter b von Patterson *et al* (vgl. [17]) $b = 1{,}018 \cdot \mathrm{ERB}$ und einer -3 dB Bandbreite von 0,887 ERB.

Die Struktur einer digitalen Gammatone-Filterbank 4. Ordnung mit N Kanälen ist in Abbildung 9 zu

4 Man kann natürlich auch direkt auf den zeitdiskreten Gammatone $h_{m,GF}[n]$ eine z-Transformation anwenden. Es stellt sich bei der Berechnung aber heraus, dass eine Faktorisierung der Nullstellen in *second order sections* auf diese Weise ungleich schwerer zu finden ist.

sehen.

Mit dem *Gain*-Faktor $g_{M,GF}$ kann der Filter so eingestellt werden, dass bei der Filtermittenfrequenz der Frequenzgang auf *unity gain* normiert ist. Eine einfache Abschätzung, die empirisch gefunden wurde und für mehrere Ordnungen stimmt ist im Anhang erklärt:

$$g_{M,GF} = \left[\frac{g_{1,APGF}}{\left(2-2\cdot\cos(\theta)\right)^{\frac{1}{5}} \left(2+2\cdot\cos(\theta)\right)^{\frac{1}{4}}} \right]^{M}, \tag{15}$$

mit dem Gain-Faktor eines Allpolfilters (oder 1. Ordnung Gammatone):

$$g_{1,APGF} = (1-r)\sqrt{1-2r\cdot\cos(2\theta)+r^2}. \tag{16}$$

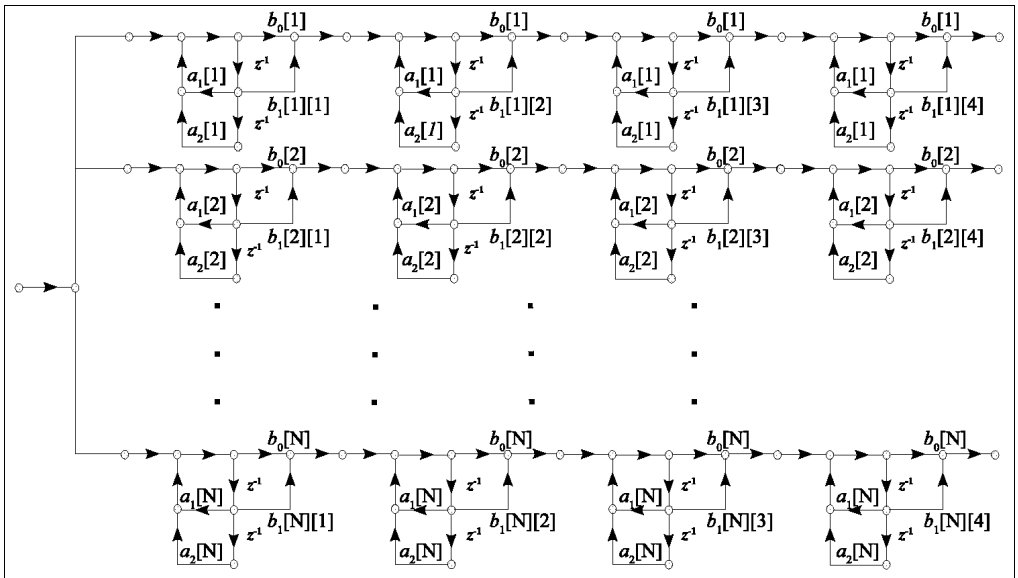

Abbildung 9: N-kanalige Gammatone-Filterbank 4. Ordnung

Für eine exaktere Lösung könnte man natürlich auch die Polynome in $z = e^{j\omega}$ an der Stelle $\omega = \theta$ auswerten.Eine N-kanalige Gammatone-Filterbank M-ter Ordnung benötigt 3MN Additionen, 4MN Multiplikationen, 2MN Speicherstellen, (3N+MN) unterschiedliche Koeffizienten, siehe Abbildung 9.

In der eigenen Simulation mit der Vorgabe der *roex*(*p,r*)-Kurven von Rosen und Baker (vgl Pflüger [7]), der *Bark*-Skala von Traunmüller (vgl Pflüger [6]) und 1 *Bark* Filterbreite, wurde die beste Übereinstimmung bei einer Bandgrenze von -8 dB gefunden, siehe Abbildung 10. Die Flanken des Filters 3. Ordnung weichen bei gleicher Bandbreite im Bereich großer Dämpfungen stärker von der vorgegebenen Form ab.

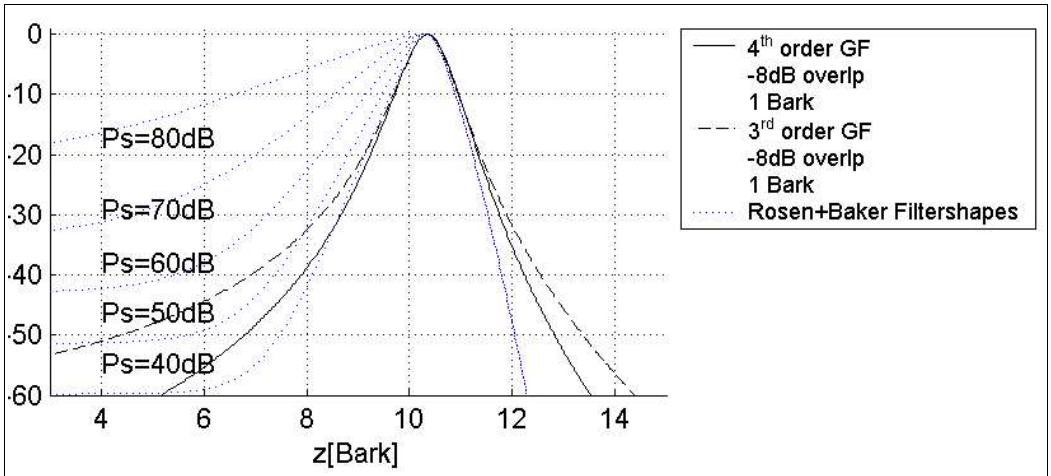

Abbildung 10: 3. und 4. Ordnung Gammatone-Filter mit 1 Bark -8dB-Bandbreite im Vergleich zu auditiven Filterformen für unterschiedliche Pegel.

3.2.3 Lineare All-Pol Gammatone-Filter (APGF)

Um All-Pol Gammatone-Filter zu erhalten werden alle Nullstellen der Laplace-/ oder *z*-Domäne zur Optimierung des Hardwareaufwandes weggelassen [17]. Vergleichend kommt die in der Arbeit von Lyon zur Kochlea Hydrodynamik gezeigte Struktur [14] mit einer All-Pol Beschreibung aus. Der APGF im Laplace-Bereich ist [15]:

$$H_{M,APGF}(s)=\frac{g_{M,APGF}}{\left[(s+b)^2+\omega^2\right]^M}, \tag{17}$$

oder im *z*-Bereich [15]:

$$H_{M,APGF}(z)=\frac{g_{M,APGF}}{\left[1-2r\cos(\theta)z^{-1}+r^2z^{-2}\right]^M}. \tag{18}$$

In Abbildung 11 sind die Pol- und Nullstellen dieser z-Transformierten in der komplexen Zahlenebene zu sehen.

Die etwas kompliziertere Impulsantwort ließe sich aus der M-fachen Faltung der 1. Ordnung All-Pol Impulsantwort mit sich selbst beschreiben:

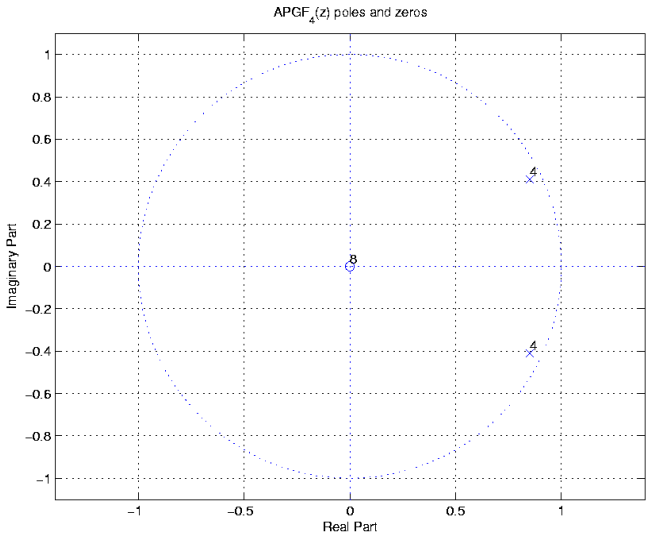

Abbildung 11: Pol-Nullstellen-Diagramm eines APGF 4. Ordnung

$$h_{1,APGF}(t) = g_{1,APGF}^{M} \cdot \frac{e^{-bt}}{\omega} \sin(\omega t) \cdot u(t), \tag{19}$$

$$h_{M,APGF}(t) = g_{M,APGF} \underbrace{h_{1,APGF}(t) * h_{1,APGF}(t) * \quad \cdots \quad * h_{1,APGF}(t)}_{M}$$

$$= g_{M,APGF} \frac{e^{-bt}}{\omega^{M}} \int_{0}^{t} \sin\left[\omega(t-\tau_1)\right] \cdot \left\{ \int_{0}^{\tau_1} \sin\left[\omega(\tau_1-\tau_2)\right] \cdot \left\{ \cdots \left\{ \int_{0}^{\tau_{M-2}} \sin\left[\omega(\tau_{M-2}-\tau_{M-1})\right] \sin(\omega\tau_{M-1})\, d\tau_{M-1} \right\} \cdots \right\} d\tau_2 \right\} d \tag{20}$$

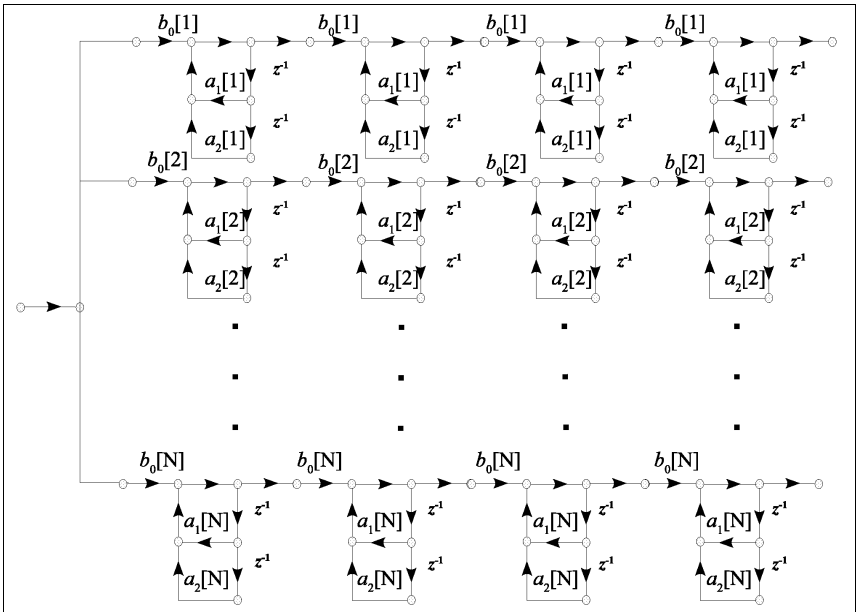

Abbildung 12: 4. Ordnung All-Pol Gammatone-Filterbank mit N Kanälen

Mit dem *Gain*-Faktor $g_{M,APGF}$ kann der APGF bei seiner Mittenfrequenz wieder auf *unity gain*

normiert werden. Hier ist die einfache Form bereits die exakte Lösung (Berechnung im Anhang):

$$g_{M,APGF} = \left(g_{1,APGF}\right)^{M}, \tag{21}$$

mit dem Gain-Faktor eines Allpolfilters (oder 1. Ordnung Gammatone) aus Gleichung 13.

Laut Lyon [15] entsteht bei dieser Faltung eine sinusförmige Trägerschwingung und eine aus Besselfunktionen. Die Berechnung der Koeffizienten, Bandbreiten und Überlappungen sind dem Anhang zu entnehmen.

Die Laplace- oder z-Bereichsbeschreibung enthält hier nur mehr M konjugiert komplexe Polpaare, wie auch in der Abbildung 12 an den rein rekursiven Strukturen zu erkennen ist. Daher braucht die N-kanalige All-Pol Gammatone-Filterbank der Ordnung M nurmehr 2MN Additionen, 3MN Multiplikationen, 2MN Speicher und 3N unterschiedliche Koeffizienten.

Wird wieder eine 1 *Bark* Bandbreite gewählt, findet sich die beste Übereinstimmung mit den auditiven Filterformen bei der Bandgrenze von -8 dB, siehe Abbildung 13. Die Flanken des Filters 3. Ordnung derselben Bandbreite weichen bei großen Dämpfungen stärker von der vorgegebenen Form ab.

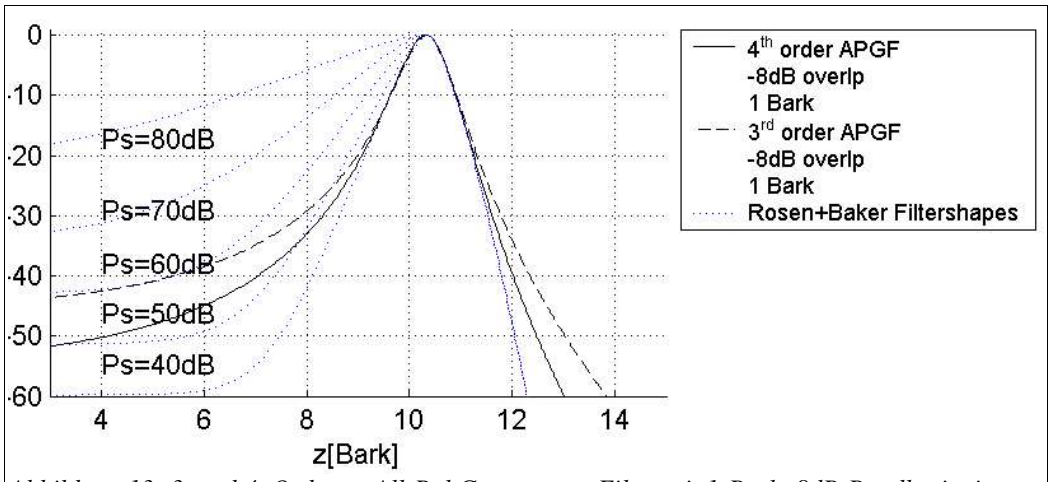

Abbildung 13: 3. und 4. Ordnung All-Pol Gammatone-Filter mit 1 Bark -8dB-Bandbreite im Vergleich zu auditiven Filterformen für unterschiedliche Pegel.

3.2.4 Lineare One-Zero Gammatone-Filter (OZGF)

One-Zero Gammatone-Filter erhält man indem man der Übertragungsfunktion eines digitalen All-Pol Gammatone-Filters eine Nullstelle bei $z = 1$ hinzufügt, oder in der Laplace-Domäne eine Nullstelle bei $s = 0$ hinzufügt [15]. Diese Hochpasscharakteristik bewirkt eine stärkere Dämpfung sehr tiefer Frequenzen, die bei der All-Pol Implementierung ohnehin zu stark durchgelassen werden.

Eine analytische Beschreibung der Impulsantwort kann durch Differenzieren der All-Pol Gammatone-Impulsantwort gefunden werden. Die Übertragungsfunktion des OZGF lautet [15]:

$$H_{M,OZGF}(s) = g_{M,OZGF} \cdot \frac{s}{\left[(s+b)^2+\omega^2\right]^M}, \tag{22}$$

$$H_{M,OZGF}(z) = g_{M,OZGF} \cdot \frac{1-z^{-1}}{\left[1-2r\cos(\theta)z^{-1}+r^2 z^{-2}\right]^M}. \tag{23}$$

In Abbildung 14 findet sich das zugehörige Pol-Nullstellen-Diagramm in der komplexen Zahlenebene.

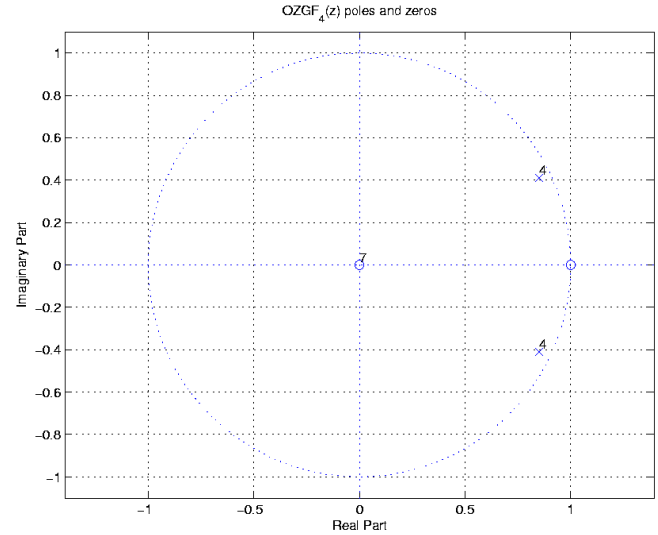

Abbildung 14: Pol-Nullstellen-Diagramm eines OZGF 4. Ordnung

Die entsprechende digitale Struktur einer N-kanaligen One-Zero Gammatone-Filterbank der Ordnung M ist in Abbildung 15 zu sehen, sie unterscheidet sich nur in der einen Nullstelle (vor der Verzweigung auf die Einzelkanäle) von der All-Pol Gammatone-Filterbank.

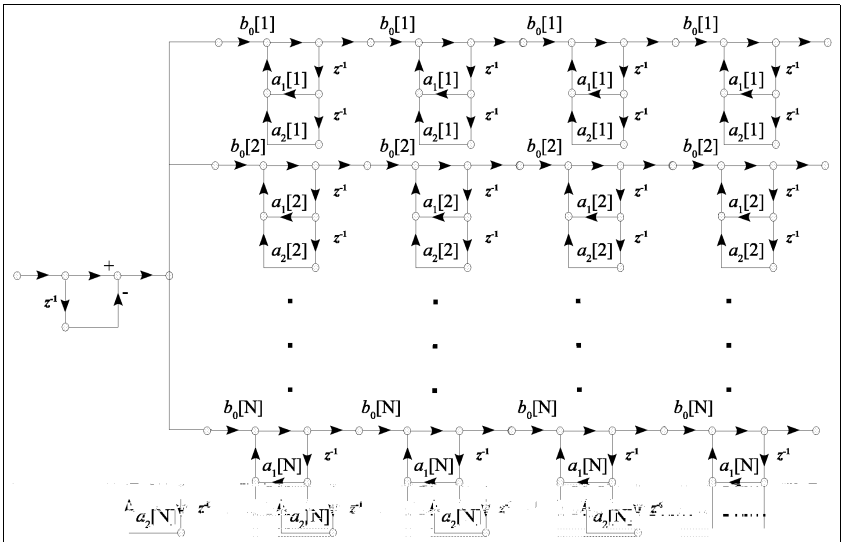

Abbildung 15: N-kanalige One-Zero Gammatone-Filterbank

Der *Gain*-Faktor für die One-Zero Gammatone Filter kann wieder aus dem Gain eines APGF (Gleichung 12) abgeschätzt werden zu (Berechnung im Anhang):

$$g_{M,OZGF} = \left[\frac{g_{1,APGF}}{\left(\sqrt{(1-\cos(\theta))^2 + \sin^2(\theta)} \right)^{\frac{1}{M}}} \right]^M . \tag{24}$$

Die N-kanalige One-Zero Gammatone-Filterbank der Ordnung M benötigt (2MN+1) Additionen, 3MN Multiplikationen, (2MN+1) Speicher und 3N unterschiedliche Koeffizienten.

Für 1 *Bark* breite Filter ergibt sich auch hier wieder eine Dämpfung von -8 dB Bandgrenze, siehe Abbildung 16.

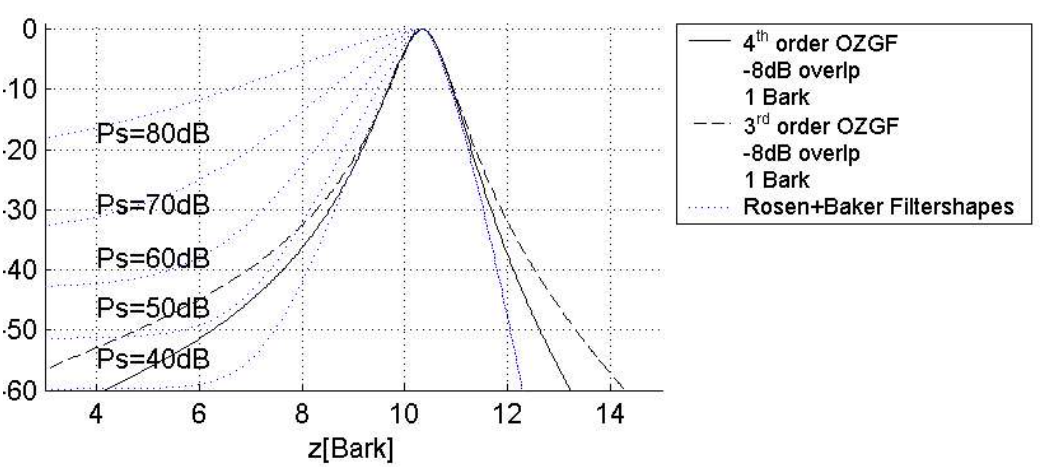

Abbildung 16: 3. und 4. Ordnung One-Zero Gammatone Filter mit 1 Bark -8dB-Bandbreite.

3.2.5 Lineare „Three-Zero“ Gammatone-Filter mit verbesserter Flanke zu hohen Frequenzen („TZGF“)

Eine Modifikation der One-Zero Gammatone-Filter kann in der Arbeit [13] von Lin, Ambikairajah und Holmes gefunden werden, in welcher ein Nullstellenpaar über der Filtermittenfrequenz verwendet wird, um die Sperrdämpfung der Gammatone-Filter zu hohen Frequenzen hin zu erhöhen und besser den gemessenen *tuning curves* anzupassen. Im z-Bereich sieht die Übertragungsfunktion dann so aus:

$$H_{M,TZGF}(z) = g_{M,TZGF} \cdot (1 - r_0 z^{-1}) \frac{1 - 2 r_z \cos(\theta_z) z^{-1} + r_z^2 z^{-2}}{\left[1 - 2 r \cos(\theta) z^{-1} + r^2 z^{-2}\right]^M}. \tag{25}$$

Für die Position der Nullstellen des zeitdiskreten Filters wurde der empirisch gefundene Zusammenhang gegeben:

$$r_0 = 0{,}955, \quad r_z = 0{,}985,$$
$$f_z = 117{,}5 (f_c/1000)^2 + 1135{,}5 (f_c/1000) + 277{,}0, \quad f_c = \frac{\omega}{2\theta} f_s, \quad \theta_z = 2\pi \frac{f_z}{f_s}. \tag{26}$$

Dabei soll das Nullstellenpaar in etwa eine Dämpfung von -60 dB ermöglichen.

Das Pol-Nullstellen-Diagramm des TZGF 4. Ordnung ist in Abbildung 11 zu sehen.

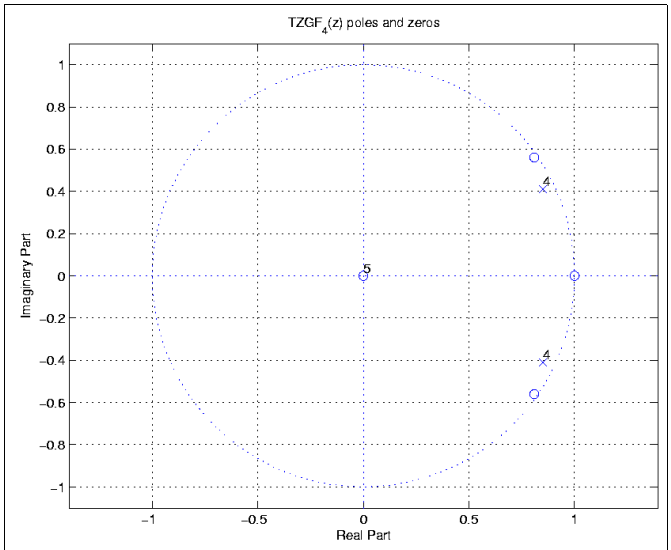

Abbildung 17: Pol-Nullstellen-Diagramm des TZGF 4. Ordnung

Nachdem die Flanke der Verdeckung zu tiefen Frequenzen etwa mit 27 dB/Bark fällt, kann man den Zusammenhang für die CB-Skala (Bark) adaptieren, indem man die Nullstellenfrequenz um 2 Bark über der Filtermittenfrequenz liegend festlegt:

$$f_z = Tfm_{CB}^{-1}\left\{Tfm_{CB}\left\{f_c\right\} + 2\,Bark\right\}, \tag{27}$$

Es muss darauf geachtet werden, dass die Nullstellen nicht bei Alias-Frequenzen zu liegen kommen, man kann die Frequenz z.B. mit f_s /2 begrenzen.

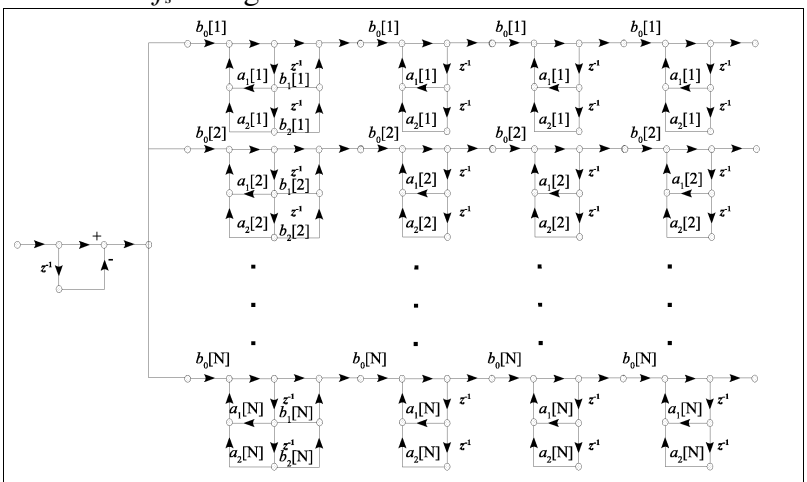

Abbildung 18: N-kanalige Three-Zero Gammatone-Filterbank

Der benötigte *Gain*-Faktor in den rein rekursiven Filterstufen entspricht dabei jenem des OZGFs, die biquadratische Filtersektion mit dem Nullstellenpaar sollte separat normiert werden, eine grobe Abschätzung ist hier schwieriger, weil die *Bark*-Transformation im Zusammenhang enthalten ist.

Deshalb wird empfohlen in die Polynome für $z=e^{j\omega}$ an der Stelle $\omega = \theta$ zur Berechnung des *Gains* zu verwenden.

Eine so berechnete Filterbank (TZGFB) würde (2MN+2N+1) Additionen, (3MN+2N) Multiplikationen, (2MN+1) Speicher und 5N unterschiedliche Koeffizienten benötigen. Im Vergleich benötigt die OZGFB M-ter Ordung gleichviele Operationen wie die TZGFB (M-1)-ter Ordnung.

Für 1 *Bark* breite Filter ergibt sich auch hier wieder eine Dämpfung von -8 dB Bandgrenze, siehe Abbildung 19. Mit dieser modifizierten Filterform kann bereits bei niedriger Ordnung eine gute Übereinstimmung mit den auditiven Filterformen erzielt werden. Der TZGF wird als geeignetste Variante von linearen Gammatone-Filtern angesehen.

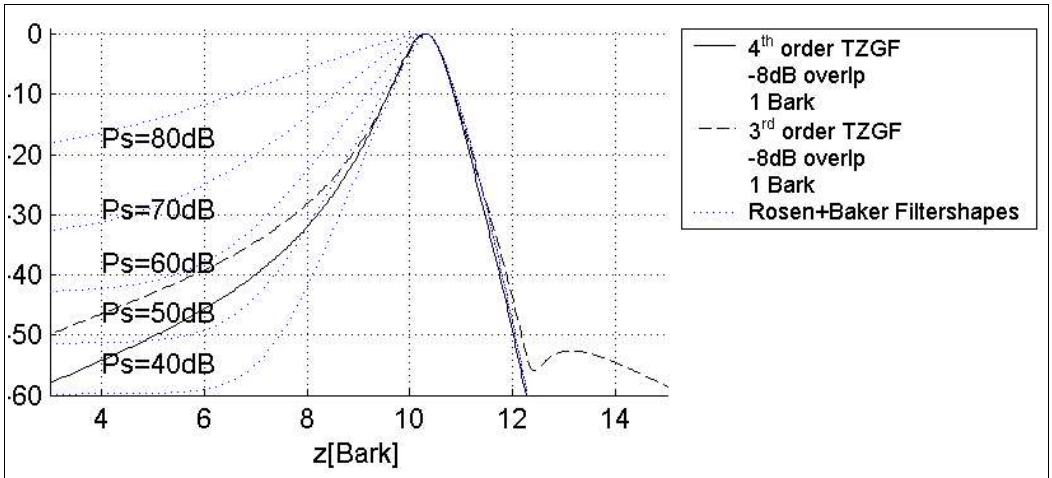

Abbildung 19: 3. und 4. Ordnung „Three-Zero" Gammatone Filter mit 1 Bark -8dB Bandbreite

3.2.6 Gammachirp, nichtlineare (pegelabhängige) Gammatone-Filter

Toshio Irino und Roy D. Patterson verwenden die Gammachirp-Funktion zur Modellierung von auditiven Filtern [44] . Die Impulsantworten dieser Filter besitzen wieder eine Gammafunktion als Amplitudeneinhüllende, aber einen Sweep oder Chirp als Trägerschwingung. Durch die Variation der Chirp-Parameter lässt sich der Filter besser den auditivem Filterformen anpassen. Grundsätzlich wäre die Implementierung einer Gammachirp-Filterbank eine aufwändige FIR-Lösung (*gammachirp wavelets*), es gibt aber die effizientere Möglichkeit über einen entsprechenden biquadratischen (*second order section*) IIR-Asymmetriefilter die IIR-Gammatone Filter in Gammachirp Filter überzuführen (*composite gammachirp* [22]). Die Abhängigkeit des Gammachirp Parameters *c* vom

Signalpegel ist bereits in der in der frühen Arbeit [21] angegeben, diese wurde aber in einer späteren Arbeit korrigiert [44]. Es hat sich herausgestellt, dass der Chirp-Parameter *c* selbst pegelunabhängig ist und die Pegelabhängigkeit mit anderen Parametern gesteuert werden muss (*compressive gammachirp*). Aus Gründen der Recheneffizienz wird diese Variante hier nicht näher betrachtet.
In Pflüger [7] werden nichtlineare, dynamisch gesteuerte APGF und OZGF vorgestellt, welche ihre Maskierungseigenschaften in Abhängigkeit des Schallpegels verändern können, wie es auch im menschlichen Gehör passiert. Von einer Implementierung wird ebenfalls aus Gründen der Recheneffizienz abgesehen, obwohl dieser Ansatz bereits relativ niedrige Komplexität besitzt.

3.2.7 Effiziente Implementierung einer inversen Gammatone-Filterbank

Um Signale, die mit einer Gammatone-Filterbank analysiert wurden, wieder zusammenzusetzen, sind einige Möglichkeiten zur Resynthese in den folgenden Abschnitten erwähnt. Insbesonders wird zum Zwecke der effizienten und latenzarmen Umsetzung den einfachen Methoden Aufmerksamkeit geschenkt.

3.2.7.1 Summenbildung

Als Synthesefilterbank stehen mehrere Varianten zur Verfügung. In der Arbet von Slaney *et al.* zur Inversion eines auditiven Modells [18] wird eine Summenbildung über alle Kanäle als einfachste Variante vorgeschlagen. Die Filterbank weist stark ansteigende Gruppenlaufzeiten zu tiefen Filter-Mittenfrequenzen hin auf, was sich in der Resynthese als Phasenverzerrung bemerkbar macht. Diese werden vom Gehör aber nicht als störend empfunden [38].

3.2.7.2 Summenbildung mit alternierendem Vorzeichen

Um Auslöschungen an den Überlappungsfrequenzen zu verhindern, sollte je nach Filterbandbreite und Ordnung zwischen benachbarten Filterbankkanälen ein Vorzeichenwechsel vor der Addition stattfinden. Der Phasenunterschied benachbarter Kanäle an ihrer Bandgrenze kann über den Phasengang eines All-Pol Filters der Ordnung M beschrieben werden. Die Berechnung dazu kann dem Anhang entnommen werden.

$$\Delta \measuredangle [H_{M,APGF}(s)] = 2M \cdot \arctan\left[\sqrt{10^{\frac{|overlap\ dB|}{10 \cdot M}} - 1}\right]. \tag{28}$$

Liegt dieser Winkel bei $(k+½)\cdot 2\pi$ mit ganzahligen Werten für k ergibt sich an der Bandgrenze eine Auslöschung, liegt der Winkel bei $k\cdot 2\pi$ ergeben sich konstruktive Überlagerungen. Wann nun ein

Vorzeichenwechsel durchzuführen ist, um eine konstruktive Überlagerung zu begünstigen, kann mit dem Zusammenhang

$$f_{sign} = sign\left(\cos\left[\Delta \measuredangle\left(H_{M,APGF}(s)\right)\right]\right). \tag{29}$$

abgeschätzt werden.

3.2.7.3 Summenbildung mit verzögerten Signalen

Sollen die Phasenverzerrungen leicht kompensiert werden, bietet sich ein Ausgleich der Gruppenlaufzeiten an den Bandgrenzen benachbarter Filter in Form von entsprechenden Laufzeitgliedern an. Wird dieser Ausgleich durchgeführt, darf kein Vorzeichenwechsel nach oben gegebener Formel erfolgen. Um eine berichtigte Abschätzung für den Phasensprung an der Bandgrenze zu erhalten, müssen unterschiedliche Verzögerungen benachbarter Bänder miteinbezogen werden:

$$\Delta \measuredangle\left[H_{M,APGF}(s)\right] = 2M \cdot \arctan\left[\sqrt{10^{\frac{|overlap\ dB|}{10 \cdot M}} - 1}\right] - \theta_{overlap} \cdot \Delta n_0, \tag{30}$$

wobei Δn_0 der Verzögerungsunterschied der Kanäle sein soll.

3.2.7.4 Verbesserte Methoden zur Resynthese

Zahlreiche andere Techniken, die zum Beispiel in den Arbeiten von Irino [23] mit zeitgespiegelten Gammatone-Filtern (ähnlich einer Wavelet-Transformation), Slaney *et al* [18] mit einer konvexen Projektion, von Lin *et al* [12] mit einem Set von optimalen Resynthesefiltern und von Kubin und Kleijn [16] mit blockweiser Zeitspiegelfilterung arbeiten, werden aufgrund des Rechenaufwands und der Latenzzeit hier nicht weiter behandelt.

3.2.8 Effiziente Implementierung einer Gammatone-Filterbank

Folgende Gesichtspunkte sind im Zuge einer latenzarmen und effizienten Implementierung zu beachten:

größere Polgüten		***größere Kanalanzahl***		***größere Filterordnung***	
pro	contra	pro	contra	pro	contra
meist bessere auditive Filterform	große Gruppenlaufzeit	meist bessere auditive Filterform	erhöhter Rechenaufwand	meist bessere auditive Filterform	drastisch erhöhter Rechenaufwand

größere Polgüten		***größere Kanalanzahl***		***größere Filterordnung***	
weniger Interferenzen bei der Resynthese	mehr Welligkeit bei der Resynthese	genaueres Modell der Frequenzgruppen		beinahe gleiche Gruppenlaufzeit bei konstanter Filterbreite	
	numerische Fehler				

Eignung der Gammatone-Implementierungsformen:

	Rechenaufwand	***Filterform***
GF: Gammatone Filter	größter	schlechteste, zu symmetrisch
APGF: All-Pole Gammatone Filter	kleinster	gut, zu großer tieffrequenter Anteil (in 8 nicht sichtbar!)
OZGF: One-Zero Gammatone Filter	gering	gut, eher symmetrisch
TZGF: „Three-Zero“ Gammatone Filter	mittel	sehr gut, Reduktion der Filterordnung möglich

Der „Three-Zero“ Gammatone Filter bietet selbst bei Verringerung der Ordnung die beste Übereinstimmung mit den auditiven Filterformen [13] bei den Lautstärkepegeln der Sprache und ist somit auch am ehesten für eine effiziente Implementierung geeignet. Die Filterform einzelner Gammaton-Filtertypen ist zu sehen in den Abbildungen 18, 8, 16 und 19.

Zur Ersparnis von Rechenleistung sollten vor allem möglichst geringe Filterordnungen, und in weiterer Folge eine geringe Kanalanzahl verwendet werden. Eine geringe Latenzzeit kann nur durch möglichst breitbandige Filter erreicht werden. Die psychoakustischen Ansprüche erfordern hingegen meist geringe Bandbreite und hohe Filterordnung.

3.2.8.1 Latenzzeit - Gruppenlaufzeit

Die Analyse mittels Gammatone-Filterbank bewirkt wie auch andere Signalanaysen eine bestimmte Laufzeit. Anders als bei der FFT-Analyse ist diese Laufzeit hier nicht für alle Frequenzen dieselbe, da sie sich aus der Gruppenlaufzeit der Filterkanäle ergibt.

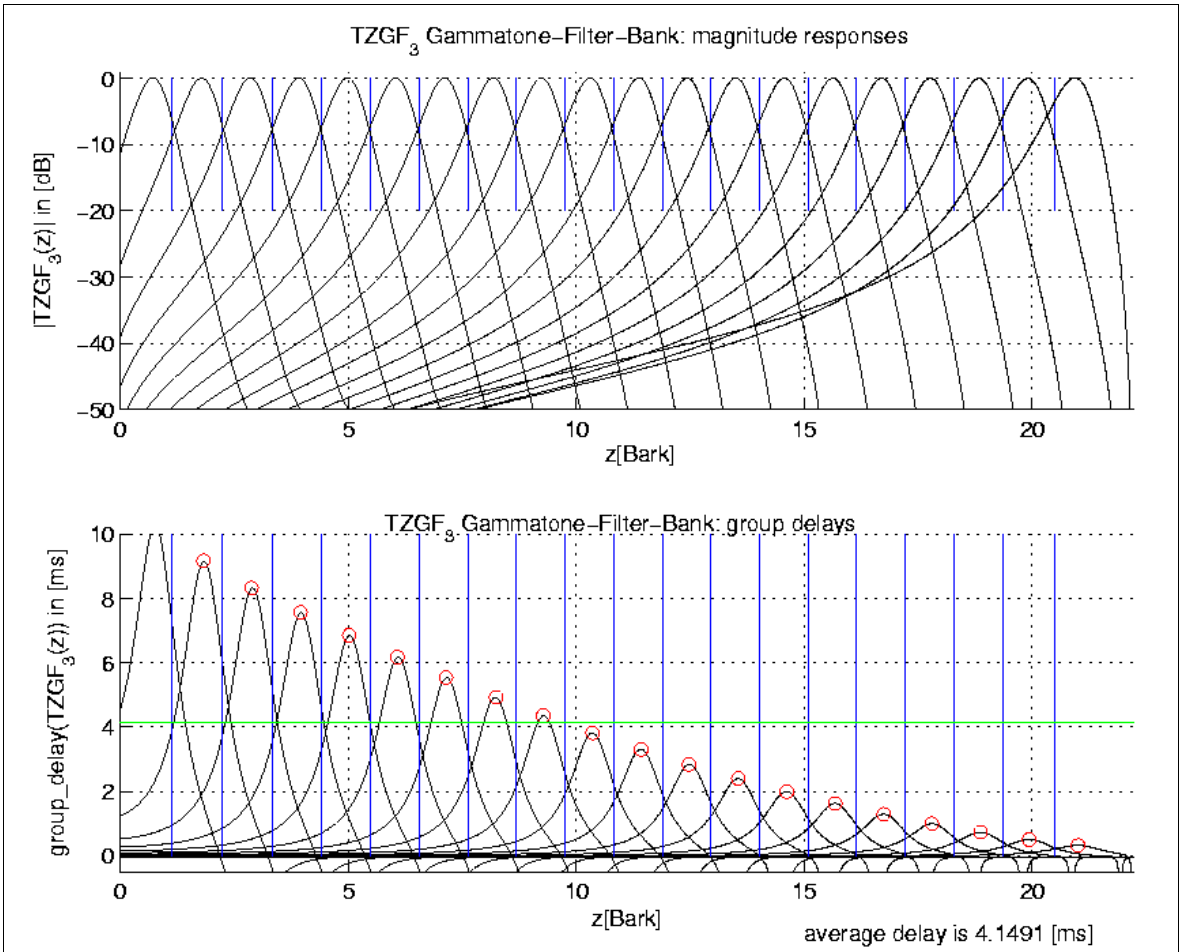

Abbildung 20 TZGFB 3. Ordnung mit -8 dB Überlappungsamplitude. Als Maß für die Latenzzeit ergibt sich aus den gemittelten Maxima der Gruppenlaufzeiten 4.15 ms

Man kann in den Simulationsergebnissen erkennen, dass sich die Filterordnung und Güte der Gammatone-Filter (von der Überlappungsamplitude abhängig) auf die Gruppenlaufzeit der Filterbank-Analyse auswirken. Es zeigt sich, dass eine Erhöhung der Filterordnung geringfügig mehr Gruppenlaufzeiten (bei konstanter Bandbreite) bewirkt. Eine Erhöhung der Güte (kleinere Überlappungsamplitude) hingegen führt zu einer starken Vergrößerung der Gruppenlaufzeit.

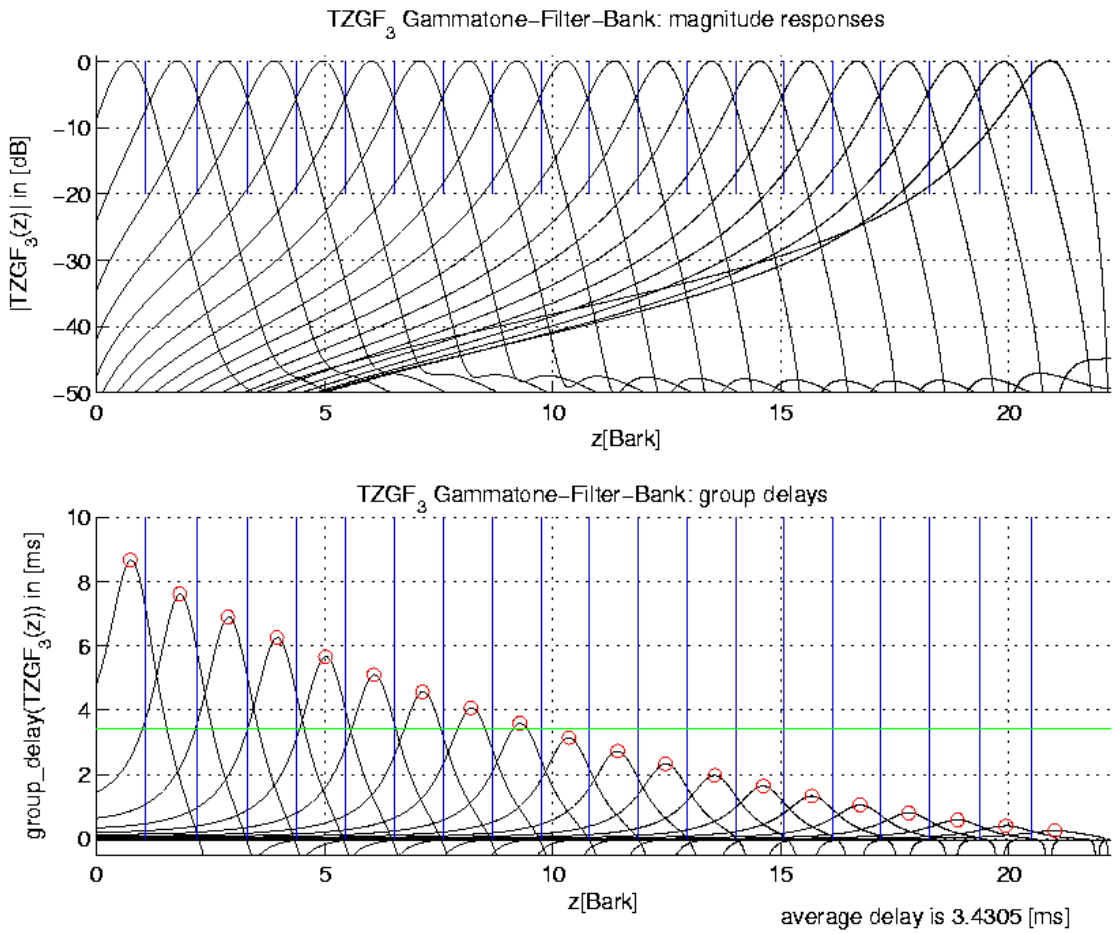

Abbildung 21: „Three-Zero“ Gammatone Filterbank (TZGFB) 3. Ordnung mit -6 dB Überlappungsamplitude. Die Durchlasskurven sind oben, die Gruppenlaufzeiten unterhalb abgebildet. Im Mittel der maximalen Gruppenlaufzeiten ergibt sich 3.43 ms als Maß für die Latenzzeit.

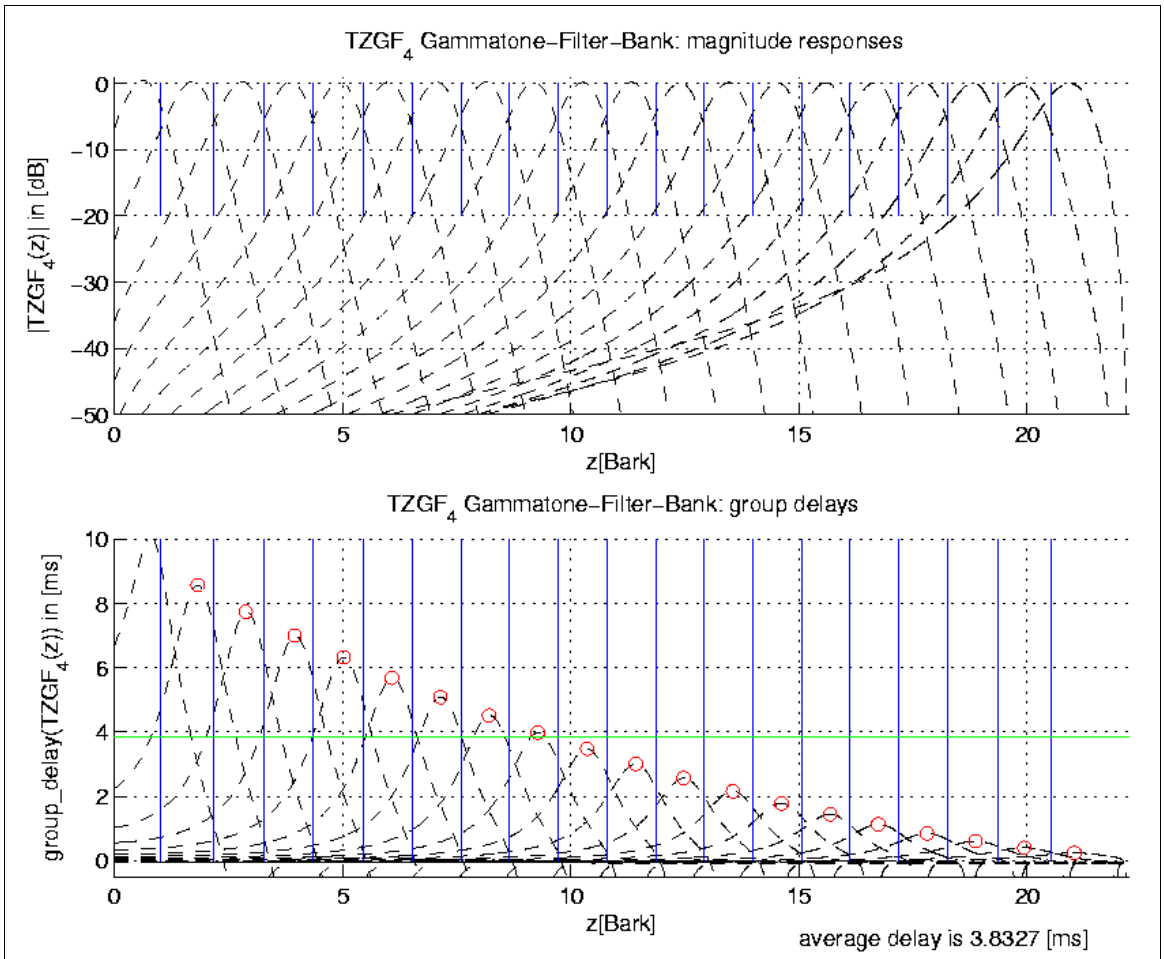

Abbildung 22: TZGFB 4. Ordnung mit -6 dB Überlappungsamplitude. Die Durchlasskurven sind oben, die Gruppenlaufzeiten unterhalb abgebildet. Im Mittel der maximalen Gruppenlaufzeiten ergibt sich 3.83 ms als Maß für die Latenzzeit.

Vergleich zu Fast-Fourier-Transform (FFT)-Analyse mit geringer Punkteanzahl:

Eine FFT-Analyse mit einer ähnlicher Anzahl von Frequenzbändern müsste etwa 64 Punkte haben. Die 32 gleichmäßig aufgeteilten Frequenzbänder dieser FFT hätten dann etwa 345 Hz oder 250 Hz Bandbreite bei Abtastfrequenzen von 22,05 kHz oder 16 kHz. Dabei hat die FFT-Analyse plus Resynthese etwa eine Latenzzeit von 1,5*N*-2*N*, also 4,3- 8 ms, wobei im Frequenzbereich unterhalb von 1,5 kHz nicht einmal annähernd eine ausreichende Frequenzauflösung gegeben ist.

Der Ansatz mit der Gammatone-Filterbank benötigt im Gegensatz zur FFT etwas mehr Rechenaufwand, liefert jedoch gleich eine dem Gehör entsprechende Frequenzauflösung samt den simultanen Maskierungseffekten des Gehörs. Die Latenzzeit bei einer Resynthese durch Addieren ist

jedoch nicht wie bei der FFT für alle Frequenzbänder gleich, da tiefe Frequenzen stärker verzögert werden als hohe. Somit ist auch keine perfekte Signalrekonstruktion möglich. Bildet man den Mittelwert der Maxima aus den Gruppenlaufzeiten der Einzelfilter (ca. bei der jeweiligen Mittenfrequenz), erhält man eine durchschnittliche Gruppenlaufzeit von etwa 3-4,5 ms. Dieser Wert ist in geringem Maße von der Filterordnung, und stark von der Filtergüte abhängig, ist aber tendenziell kleiner als jene Latenzzeit der FFT-Analyse und Resynthese.

3.2.8.2 Mithörschwelle bei der Signalanalyse mittels Gammatone-Filterbank

In diesem Abschnitt soll grafisch veranschaulicht werden, dass die Signalamplituden einer auditiven Filterbank durch die spezielle überlappende Filterform die Mithörschwelle ergeben. Auf die selbe Weise ergibt sich in der Signalamplitude die Maskierung leiser Signale, siehe Abbildung 17.

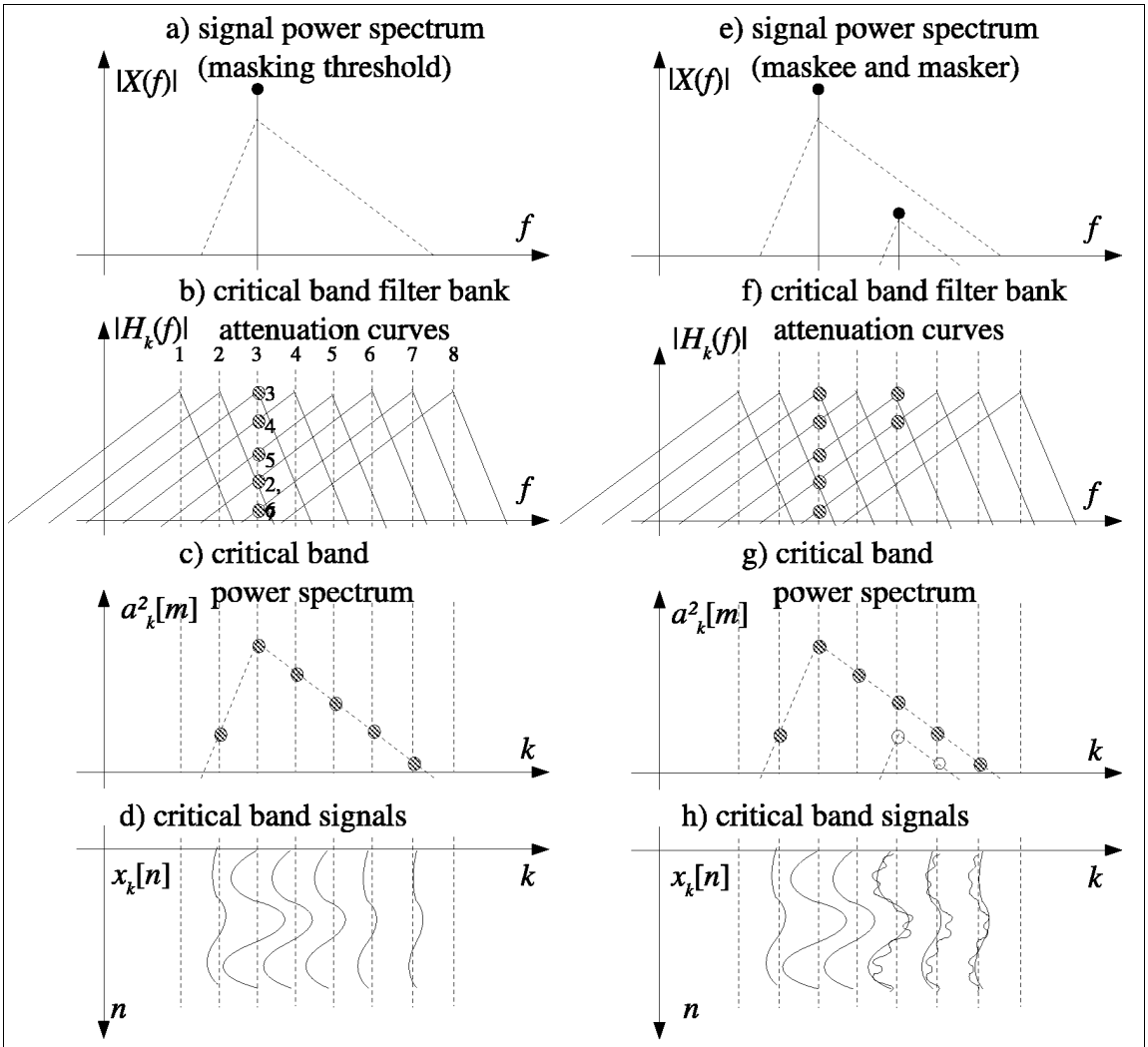

Abbildung 23: a)-d): Mithörschwelle in den Amplituden einer Gammatone-Filterbank mit kritischen Bandbreiten. e)-h): Simultanverdeckung in den Amplituden einer Gammatone-Filterbank mit kritischen Bandbreiten.

Man sieht, dass die Filterbank-Kanäle so ineinander übersprechen, dass die Kanalamplituden die Verdeckungseffekte beinhalten.

3.2.8.3 Möglichkeiten zur Blockfaltung mittels Gammatone-Filterbank

Mit der Gammatone-Filterbank ist auch die Durchführung einer Blockfilterung möglich, geht man von einer perfekten Rekonstruktion aus (die allerdings schwer implementierbar ist), entspricht die Blockfilterung einer komplexen Gewichtung der einzelnen Filterbankkanäle, wobei dazu eine Analyse mit *In-Phase* (Realteil) und *Quadratur* (Imaginärteil) Komponente durchzuführen ist. Bei der Blockfaltung mittels FFT sind Faltungen mit großen Impulsantwortlängen, welche die FFT-Punkteanzahl übersteigen, dann möglich, wenn mehrere (entsprechend der FFT-Fensterlänge)

verzögerte FFT-Analysen mit FFT-analysierten Teilen der Impulsantwort blockgefaltet werden. Auch bei der Blockfaltung mit der Gammatone-Filterbank ist es möglich längere Impulsantworten zu verwenden. Die Fensterlänge der Einzelkanäle ist aber bei tiefen Frequenzen sehr lang (ca. 8 ms), und bei hohen Frequenzen sehr kurz (ca. 0,2 ms). Um für alle Frequenzen gleich lange Impulsantworten zu ermöglichen, müsste man entsprechend der zeitlichen Ausdehnung der Kanalimpulsantworten (=Fensterlänge) verzögerte Versionen der Bandpassignale komplex gewichtet aufaddieren.

3.3 Erzeugung von In-Phase und Quadratur Signalkomponenten

Zum unveränderten *In-Phase* Signal eines Systems kann über einen Hilbert-Transformator ein *Quadratur* Signal berechnet werden. Mit beiden Signalen zusammen erhält man ein *analytisches* Signal, wobei die *In-Phase* Komponente den Realteil und die *Quadratur* Komponente den Imaginärteil eines komplexen Signales darstellen.

Durch Summenbildung über unterschiedlich gewichtete *In-Phase* und *Quadratur* Signale kann jede beliebige Phasendrehung im ursprünglichen Signalspektrum erzielt werden, zudem ist es möglich die Amplitudeneinhüllende über die Summe der Quadrate beider Signale zu berechnen.

Diese Eigenheiten sind bei der Auslöschung von Echos (EC, *echo cancellation*) und bei der Bestimmung der Signalamplitude wertvoll.

Ein Hilbert-Transformator, der im gesamten Frequenzbereich eine 90°-Phasenverschiebung erzeugen kann, besitzt aber eine sehr aufwendige Implementierung. Bei der Verwendung von Schmalbandsignalen muss die gewünschte Phasendrehung nur in einem begrenzten Bereich des Spektrums stimmen. Methoden für schmalbandige 90° Phasenschieber werden in den folgenden Abschnitten beschrieben.

3.3.1 Verzögerungen

Eine einfache Möglichkeit zur 90° Phasendrehung in einem schmalbandigen Bereich ist der Einsatz einer Verzögerung um eine ¼ Signalperiode. Die benötigte Verzögerung dazu ist also:

$$n_0 = \frac{f_s}{4f} = \frac{\pi}{2\theta}. \tag{31}$$

Stehen nur ganzzahlige Verzögerungen zur Verfügung, beschränkt sich der Einsatzbereich auf

Frequenzen unterhalb von $\theta < \pi/2$. Es besteht die Möglichkeit des Einsatzes fraktionaler Delays um eine genauere Abstimmung zu ermöglichen.

Der Phasenfehler bei abweichender Signalfrequenz kann durch eine Linearisierung berechnet werden. Dabei wird die Steigung der Phase um die Frequenz θ durch Ableitung nach $d\omega$ gebildet, dies entspricht der Gruppenlaufzeit. (*Anm.: Ein idealer Hilbert-Transformator mit konstanter und exakter 90° Phase hätte keine Gruppenlaufzeit und wäre akausal*).

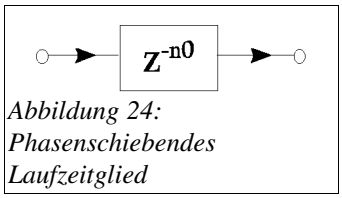

Abbildung 24: Phasenschiebendes Laufzeitglied

Durch Multiplikation mit der Frequenzabweichung $\Delta\omega = (\omega - \theta)$ ergibt sich die Näherung für den Phasenfehler $\Delta\varphi$, in weiterer Folge wird diese durch die relative Frequenzabweichung δ_θ ersetzt:

$$\Delta\varphi = \Delta\omega \cdot grd\left[e^{-j\omega n_0}\right] = \Delta\omega \cdot n_0 = \Delta\omega \cdot \frac{\pi}{2\theta},$$
$$\Delta\varphi = (\delta_\omega^{\pm 1} - 1)\theta \cdot \frac{\pi}{2\theta} = (\delta_\omega^{\pm 1} - 1)\frac{\pi}{2}. \tag{32}$$

Bei Frequenzbändern mit 1 Bark Breite mit einer Mittenfrequenz von 100 Hz beginnend, liegt die größte Abweichung bei etwa $\delta_\theta = 1{,}4$ (<verminderte Quint, Bark 1), daraus würde sich ein maximaler Phasenfehler von $\Delta\varphi = \pm\pi/5$ ergeben.

Für das tiefste Frequenzband benötigt man etwa 50 Samples Verzögerung. Wird eine ganzzahlige Verzögerung als Ringbuffer implementiert, ist diese Methode zur Realisierung eines Phasenschiebers die effizienteste und nummerisch stabilste aller 3 genannten. Der Nachteil dieses Verfahrens ist die große Phasenungenauigkeit. Rechenaufwand 0, Speicherstellen weniger als 50 Samples.

3.3.2 FIR 1. Ordnung

Aus der Gleichung für den Phasengang eines FIR-Systems 1. Ordnung kann durch Einsetzen der Forderung einer 90° Phasendrehung ($=\pi/2$) bei der Frequenz $\omega=\theta$ die reelle Nullstelle $r_{1,90°}$ gefunden werden:

$$arg[H_{1,90^\circ}(e^{j\omega})] = -\arctan\left(\frac{r_{1,90^\circ}\sin(\omega)}{1-r_{1,90^\circ}\cos(\omega)}\right) \stackrel{!}{=} -\frac{\pi}{2}\Bigg|_{\omega=\theta},$$
$$\Rightarrow r_{1,90^\circ} \stackrel{!}{=} \frac{1}{\cos(\omega)}. \tag{33}$$

Der gesuchte Nullstellenfilter besitzt somit die Übertragungsfunktion:

$$H_{1,90^\circ} = \left[1 - r_{1,90^\circ} z^{-1}\right] \cdot \frac{1}{\sin(\theta)} = \left[1 - \frac{1}{\cos(\theta)} z^{-1}\right] \cdot \frac{1}{\sin(\theta)}. \tag{34}$$

Dieser einfache maximalphasige Filter lässt sich mit gutem Erfolg zur 90° Phasenverschiebung eines Gammatone-Filters einsetzen, wobei der Frequenzgang durch eine Hochpass-Charakteristik verfälscht wird, wenn $\theta < \pi/2$, oder durch eine Tiefpass-Charakteristik wenn $\theta > \pi/2$. (*Anm.: echte Gammatone-Filter haben unabhängig von Sinus- bzw. Cosinus-Phase identische Frequenzgänge, die Nullstellen unterscheiden sich in der Anzahl <u>und</u> in den Positionen*)

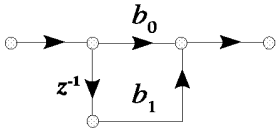

Abbildung 25:
Phasenschiebender FIR-Filter

Der von der Frequenzabweichung $\Delta\omega$ abhängige Phasenfehler kann wieder über die Gruppenlaufzeit beschrieben werden zu:

$$\Delta\varphi = \Delta\omega \cdot grd\left[1 - \frac{1}{\cos(\theta)} e^{-j\omega}\right]_{\omega=\theta<\frac{\pi}{2}} = \Delta\omega \cdot \frac{\frac{1}{\cos^2(\theta)} - \frac{1}{\cos(\theta)}\cos(\theta - 0)}{1 + \frac{1}{\cos^2(\theta)} - 2\frac{1}{\cos(\theta)}\cos(\theta - 0)} = -\Delta\omega. \tag{35}$$

Für eine relative Frequenzabweichung δ_θ umgeschrieben ergibt sich der Phasenfehler zu

$$\Delta\varphi = (\delta_\theta^{\pm 1} - 1)\theta. \tag{36}$$

In einer Filterbank mit 1 Bark-breiten Kanälen ab einer Mittenfrequenz von 100 Hz ergibt sich die maximale Frequenzabweichung in den ersten Kanälen mit dem Frequenzverhältnis $\delta_\theta = 1{,}4$ (< verminderte Quint). Es ist daher bei Verwendung dieses Verfahrens mit einem maximalen Phasenfehler von etwa $\Delta\varphi = \pm\pi/7$ zu rechnen.

Der Vorteil dieses 90° Phasenschiebers besteht in der minimalen Komplexität der Implementierung und in einer sehr guten Phasengenauigkeit. Ein Nachteil ist die Veränderung des Durchlassbereiches durch die jeweilige Filtercharakteristik und numerische Probleme, die sich aus 1/sin(θ) und 1/cos(θ) in der Nähe von θ = 0, π und θ = π/2 ergeben. Rechenaufwand 1 Addition, 2 Multiplikationen, 1 Speicherstelle, 2 Koeffizienten.

3.3.3 Allpass 1. Ordnung

Setzt man in die Gleichung für den Phasenwinkel eines Allpasses 1. Ordnung die Forderung einer 90° Phasendrehung (=π/2) bei der Frequenz ω=θ ein, kann der Koeffizient $r_{1.\mathrm{AP},90°}$ der Übertragungsfunktion $H_{1.\mathrm{AP},90°}(z)$ gefunden werden:

$$arg[H_{1.AP,90°}(e^{j\omega})] = -\omega - 2\arctan\left(\frac{r_{1.AP,90°}\sin(\omega)}{1-r_{1.AP,90°}\cos(\omega)}\right) \stackrel{!}{=} \frac{\pi}{2}\Bigg|_{\omega=\theta}, \tag{37}$$

$$\Rightarrow r_{1.AP,90°} \stackrel{!}{=} \frac{1}{\cos(\theta) - \frac{\sin(\theta)}{\tan\left(\frac{\theta}{2}+\frac{\pi}{4}\right)}},$$

$$H_{1.AP,90°}(z) = \frac{r_{1.AP,90°} - z^{-1}}{1 - r_{1.AP,90°} z^{-1}}. \tag{38}$$

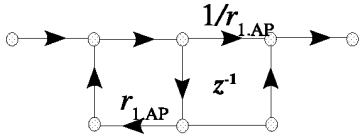

Abbildung 26: Phasenschiebender Allpass

Der Phasenfehler dieser Anordnung könnte wieder über die Gruppenlaufzeit bestimmt werden, die Berechnung dazu bringt einen unschönen Ausdruck, der zeigt, dass die Phasenfehler hier frequenzabhängig sind. In der Simulation ist ersichtlich, dass die Phasenfehler im Vergleich stärker ausgeprägt sind, als jene des oben beschriebenen FIR-Systems 1. Ordnung.

Dieser Allpass kann verwendet werden, um einen Gammatone-Filter 90° in der Phase zu drehen, ohne dabei den Frequenzgang zu beeinflussen. Der Rechenaufwand dieser Implementierung hält

sich im Vergleich zu einem breitbandigen Hilbert-Transformator auch noch in Grenzen, numerische Schwierigkeiten kann es im Bereich $\theta=0, \pi$ geben. Rechenaufwand 2 Additionen, 2 Multiplikationen, 1 Speicherstelle, 2 Koeffizienten.

3.3.4 Kombination mehrerer Allpässe

Um eine Phasendrehung in einem breiteren Bereich des Spektrums zu erreichen, können besondere Allpass-Strukturen eingesetzt werden. Solche Anordnungen können zwar nicht einzelne Signale alleine um 90° in der Phase drehen, dafür aber zwei Signale so bearbeiten, dass dazwischen eine 90° Phasendifferenz auftritt.

Zur Dimensionierung sucht man sich wieder eine Mittenfrequenz, um welche herum die beiden Allpässe der Ordnung M eine genaue 90° Phasendifferenz zueinander haben müssen. Der Phasenwinkel M 90° ist die Hälfte der gesamten erreichbaren Allpass-Phasendrehung. Von dieser Phasendrehung ausgehend setzt man einen Allpass auf M 90° + 45°, den anderen auf M 90° - 45° bei der festgelegten Mittenfrequenz. Dadurch erreicht man für diese spezielle Frequenz eine exakte 90° Phasendifferenz, für benachbarte Frequenzen eine angenäherte 90° Phase.

Diese Dimensionierung kann nun auch für mehrere Frequenzen durch Verkettung solcher Allpass-Strukturen vorgenommen werden, um eine breitere Wirkung zu erzielen.

3.4 Leistungsbestimmung

3.4.1 Nichtlinearität und Mittelung

In Abbildung 27 wird eine einfache Struktur vorgeschlagen, die durch eine nichtlineare Kennlinie (z.B. Betrag, Halbwellengleichrichtung, ...) und Tiefpassfilterung die *Amplituden-Einhüllende* eines Bandpasssignals bestimmt. Die nichtlineare Kennlinie erzeugt Summen- und Differenzfrequenzen im Spektrum des Bandpasssignals. Darin enthalten ist auch ein langsam veränderlicher Signalanteil, der die Signalamplitude beschreibt. Mit dem Tiefpassfilter wird dieser Signalanteil, der in der Umgebung des Gleichanteils (*DC*) im Spektrum gefunden wird, extrahiert und zur Beschreibung der Amplitude verwendet.

Wesentliche Voraussetzungen zur korrekten Funktionsweise sind:

1. Durch die Nichtlinearität soll kein Aliasing verursacht werden.
2. Bei mehr als einer Oktave Bandbreite können Mischfrequenzen (ω_1 - 2 ω_2) im Gleichanteil zu

liegen kommen, deshalb sollte die Signalbandbreite kleiner als eine Oktave sein.

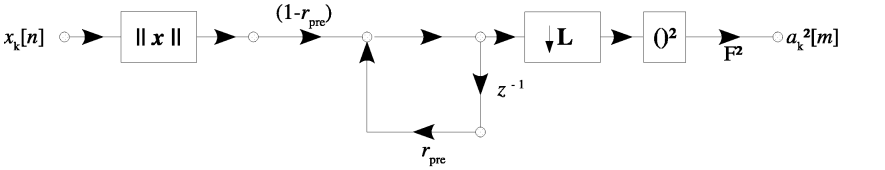

Abbildung 27: Bestimmung der Amplituden-Einhüllenden von $x_k[n]$

Werden diese Voraussetzung nicht eingehalten, können im Durchlassbereich (*DC*) des Glättungsfilters zusätzliche Signalkomponenten entstehen, welche nicht die Signalamplitude beschreiben. Bei mehreren Schwingungen treten alle Mischfrequenzen der Schwingungen auf, was im Falle von Schwebungsfrequenzen und Differenztönen den Eigenschaften des Gehörs entspricht.

Je nach Anwendung können verschiedene nichtlineare Kennlinien verwendet werden, wie zum Beispiel Quadrieren oder Logarithmieren. Am effizientesten ist vermutlich die Absolutbetragbildung, diese ist nur eine 1 Bit-Operation und erzeugt Signalkomponenten im Basisband und bei der doppelten Frequenz. Die benötigte Mittelungskonstante muss daher auch nicht sehr groß sein, um entstehende Welligkeiten zu glätten, da im ursprünglichen Frequenzband keine Mischprodukte entstehen. Die Berechnung des Amplitudenquadrats zur Leistungsbestimmung kann auch in einem unkritischen Bereich mit niedriger Samplingrate durchgeführt werden. In jedem Fall ist der Formfaktor F des gleichgerichteten Signals zu beachten, der zur ermittelten Signalamplitude hinzumultipliziert werden muss.

Halbwellengleichrichtung wird in physiologischen Modellen häufig verwendet, um den *Transduktionsprozess der inneren Haarzellen* zu modellieren [18][22], und könnte deshalb auch hier alternativ zum Absolutbetrag eingesetzt werden. Allerdings ist der benötigte Glättungsfilter hier mit einer größeren Zeitkonstante zu betreiben als bei Absolutbetragsbildung.

Der Vorteil dieses Modells zur Pegelbestimmung besteht in der Einfachheit der Implementierung. Ein Gleichrichter ist im Wesentlichen eine einfache Bit-Operation, die das Vorzeichen-Bit löscht. Es ergeben sich für N Kanäle N Absolutbetragsoperationen, 2N+2N / L Multiplikationen und N Additionen.

Weiters ist es vorteilhaft, dass bereits hier bei der Tiefpassfilterung eine „zeitliche Verschmierung“ mit mindestens 2 ms Mittelungsdauer durchgeführt wird, welche die Vorverdeckung des Gehörs modelliert [19]. Die zugehörigen Bandpasssignale sollten dabei entsprechend verzögert werden, um

die Gruppenlaufzeiten der Mittelung auszugleichen.

Ein Nachteil dieses Verfahrens ist, dass die Mittelungsdauer bei tiefen Frequenzen hoch sein muss, um unerwünschte Modulationen der Signalamplitude zu unterdrücken. Das führt zu unerwünscht hohen Gruppenlaufzeiten bei tiefen Frequenzen, die ohnehin bereits bei der Filterung am stärksten verzögert werden.

3.4.1.1 Bandbreite des Amplitudensignals (Nichtlinearität+Mittelung)

Die maximale Bandbreite des gefundenen Amplitudensignals wird entweder durch den eingesetzten Tiefpassfilter vorgegeben, oder durch die Bandbreite des untersuchten Bandpasssignals. Für sehr tieffrequente Signale sollten längere Zeitkonstanten gewählt werden, um durch die Nichtlinearität erzeugte Signalkomponenten noch ausreichend zu unterdrücken (Welligkeit).

Für Frequenzbänder mit 1 Bark Breite und deutlich höherer Mittenfrequenz, wird eine Mittelungsdauer von 2-3 ms vorgeschlagen, was der Zeitkonstante der Vorverdeckung durch das Gehör entspricht.

Im angestrebten Verwendungszweck sind die Bandpasssignale bei tiefen Frequenzen etwa ein Bark breit, die Mittelungsdauer sollte in diesen Bändern (etwa < 1000 Hz) der halben Bandbreite entsprechen und deshalb größer als 2-3 ms sein.

Somit sind die Bandbreiten der Amplitudensignale für Bänder unter etwa 1000 Hz jeweils ein halbes Bark, für Bänder mit höherer Mittenfrequenz (> 1000 Hz) etwa 500 Hz. Um die Gruppenlaufzeiten der Amplitudenmittelung auszugleichen, wird vorgeschlagen die zugehörigen Bandpasssignale etwa um die halbe Mittelungsdauer (1 ms) zu verzögern.

3.4.2 In-Phase und Quadratur Signale zur Amplitudenbestimmung

Für harmonische Signale ist folgender Ansatz zur Bestimmung der Amplituden-Einhüllenden $a[n]$ über die *In-Phase* $x_{k,I}[n]$ und die *Quadratur* Komponente $x_{k,Q}[n]$ zulässig:

$$\begin{aligned} x_{k,I}[n] = a[n]\cdot\cos(\theta n) \quad &<==> \quad x_{k,Q}[n] = a[n]\cdot\sin(\theta n) \quad . \\ a[n] = \sqrt{x_{k,I}^2[n] + x_{k,Q}^2[n]} = \sqrt{a^2[n]}\cdot\sqrt{\sin^2(\theta n)+\cos^2(\theta n)} &= a[n] \quad . \end{aligned} \tag{39}$$

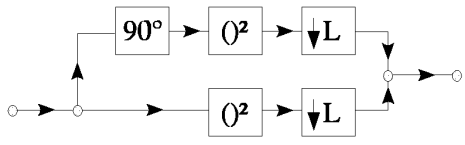

Abbildung 28: Bestimmung des Amplitudenquadrates über Phasenschieber

Werden die beiden Signalkomponenten unterschiedlicher Phase quadriert und aufsummiert, lässt sich die harmonische Schwingung herauskürzen und es bleibt die quadrierte Amplitudeneinhüllende übrig, siehe Abbildung 28.

Treten mehrere harmonische Schwingungen parallel auf ergibt sich, der Gehörwahrnehmung entsprechend, die Summe der Leistungen und die Schwebungsfrequenz, bzw. ein Rauhigkeitseffekt in der Amplitude. Das kann anhand vom Beispiel mit 2 Schwingungen gezeigt werden:

$$\begin{aligned} a_{1,2}^2[n] &= \left(a_1[n]\cdot\sin(\theta_1)+a_2[n]\cdot\sin(\theta_2)\right)^2+\left(a_1[n]\cdot\cos(\theta_1)+a_2[n]\cdot\cos(\theta_2)\right)^2, \\ &= a_1[n]^2\left(\sin^2(\theta_1)+\cos^2(\theta_1)\right)+a_2[n]^2\left(\sin^2(\theta_2)+\cos^2(\theta_2)\right)+2\,a_1[n]a_2[n]\left(\sin(\theta_1)\sin(\theta_2)+\cos(\theta_1)\cos(\theta_2)\right), \\ &= a_1^2[n]+a_2^2[n]+2\,a_1[n]a_2[n]\sin(\left|\theta_2-\theta_1\right|). \end{aligned} \tag{40}$$

Methoden, wie aus einem (*In-Phase*) Signal ein *Quadratur* Signal durch 90°-Phasendrehung erzeugt werden kann sind in einem der vorigen Abschnitte zu finden.

3.4.2.1 Auswirkung von Phasenfehlern auf das Amplitudensignal

Besteht zwischen dem *In-Phase* und dem *Quadratur* Signal nur näherungsweise eine 90°-Phasenverschiebung fallen diese Ungenauigkeiten bei der Bildung eines Amplitudenwertes auf. Der Phasenfehler $\Delta\varphi$ verursacht Modulationen in der Amplitude, die für einfache harmonische Schwingungen so aussehen:

$$\begin{aligned} \hat{a}[n] &= \sqrt{a^2[n]\sin^2(\omega t)+a^2[n]\cos^2(\omega t+\Delta\varphi)}, \\ \hat{a}[n] &= a[n]\sqrt{1+\frac{1}{4}\left[e^{j2\omega t}e^{j2\Delta\varphi}-e^{-j2\omega t}+e^{-j2\omega t}e^{-j2\Delta\varphi}-e^{j2\omega t}\right]}, \\ \hat{a}[n] &= a[n]\sqrt{1-\sin(\Delta\varphi)\cdot\sin(2\,\omega t+\Delta\varphi)}. \end{aligned} \tag{41}$$

$$w_{dB} = 10\log_{10}\left[\frac{1+\sin(\Delta\varphi)}{1-\sin(\Delta\varphi)}\right]. \tag{42}$$

Die modulierende Schwingung besitzt die doppelte Grundfrequenz der analysierten Schwingung und tritt verzerrt in radizierter Form auf. Die dadurch entstehende Welligkeit w_{dB} der Amplitude hängt direkt mit dem Phasenfehler zusammen. Würde man die Welligkeit auf $w_{\mathrm{dB}}<2$ dB begrenzen, so

müsste der Phasenfehler $\Delta\varphi < \pi/13{,}5$ sein.

3.4.2.2 Bandbreite des Amplitudensignals (In-Phase+Quadratur)

Die Bandbreite eines kontinuierlichen Amplitudensignals $a(t)$, das mit dieser Methode bestimmt wurde, entspricht etwa der halben Bandbreite des Bandpasskanals, am leichtesten ist diese über die Kanal-Impulsantwort zu ermitteln. Hier am Beispiel des Gammatones M-ter Ordnung

$$h_{M.GF}[n] = t^{M-1} e^{-bt} \cos(\omega t). \tag{43}$$

Für die Impulsantwort des Gammatone-Filters ergibt sich als Amplitudeneinhüllende $a_{GF}(t)$ die *Gammafunktion* M-ter Ordnung, durch welche die Bandbreite des Amplitudensignals bestimmt wird:

$$a_{GF}(t) = \sqrt{t^{2(M-1)} e^{-2bt} \cos^2(\omega t) + t^{2(M-1)} e^{-2bt} \sin^2(\omega t)} \quad \Rightarrow \quad a_{GF}(t) = t^{M-1} e^{-bt}. \tag{44}$$

Die Laplace-Transformierte der Gammafunktion kann in einer geschlossenen Form für alle Ordnungen angeschrieben werden zu:

$$A_{GF}(s) \;\propto\; \frac{1}{(s+b)^M}. \tag{45}$$

Aus diesem Ausdruck lässt sich Bandbreite B_A des Amplitudensignals bei gewünschte Sperrdämpfung L_{dB} an der Bandgrenze errechnen, mit $b = f_s \cdot \ln(r)$:

$$B_A = \frac{\sqrt{10^{\frac{L_{dB}}{10 \cdot M}} - b^2}}{2\pi} = \frac{\sqrt{10^{\frac{L_{dB}}{10 \cdot M}} - f_s^2 \cdot \ln^2(r)}}{2\pi}. \tag{46}$$

Ein vereinfachter Zusammenhang, der die bekannte Bandbreite B_{APGF} des Gammatone-Filters, den Pegel C_{dB} an der Bandgrenze und die ungefähre Abschätzung der Flankensteilheit 6 M dB/Okt benützt ist gegeben durch:

$$B_A \simeq \frac{B_{APGF}}{2} 2^{\frac{L_{dB} - C_{dB}}{6M}}. \tag{47}$$

Ein Vorteil dieser Methode zur Pegelbestimmung besteht darin, dass selbst bei tiefen Frequenzen keine zusätzlichen Gruppenlaufzeiten zur Glättung von Welligkeiten eingebracht werden müssen. Bei einer Gammatone-Filterbank mit Bandbreite von 1 Bark kann bis zu Mittenfrequenzen von etwa 15 Bark noch etwa 20-fach unterabgetastet werden. Es sind keine Verzögerungen nötig um die Bandpassignale mit den zugehörigen Amplitudensignalen zu synchronisieren.

Der Nachteil dabei ist, dass das genannte Verfahren für hohe Frequenzen erstens nicht die Zeitkonstanten der Vorverdeckung des menschlichen Gehörs nachbilden kann und zweitens keine ausreichende Glättung für die gewünschte Unterabtastung erreicht. Die „zeitliche Verschmierung“, bzw. eine Anti-Aliasing Fensterung müsste daher in einem separaten Schritt bei hoher Abtastfrequenz durchgeführt werden.

3.5 Zeitliche Vor- Nachmaskierung

Die zeitliche Maskierung des Gehörs beschreibt die Verdeckung in der Wahrnehmung von Schallereignissen durch zeitliche Nähe zu anderen, meist lauteren, Schallereignissen. Grundsätzlich können 2 Arten von zeitlicher Maskierung beobachtet werden (Abbildung 33 [20]).

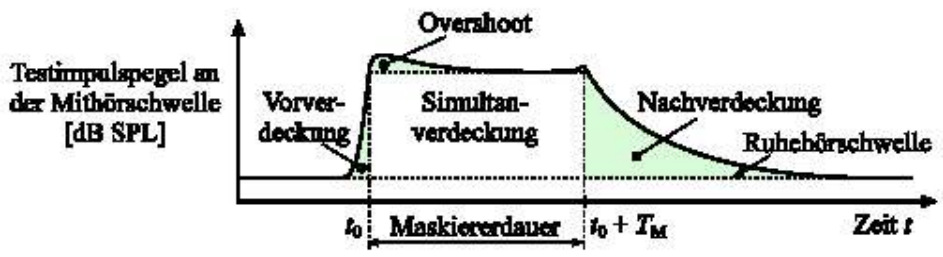

Abbildung 29: Vor- und Nachverdeckung des Gehörs (Frank Baumgarte [20]*)*

3.5.1 Modell für die Vormaskierung

Die Vorverdeckung ist für alle Frequenzbänder etwa gleich, ist aber nicht für jeden Menschen gleich ausgeprägt. In der Dissertation von Frank Baumgarte [20] wird zur zeitlichen Verschmierung bzw. Vorverdeckung ein FIR-Filter mit einem cos²-förmigen Impulsantwort verwendet. Diese Form kann auch in PEAQ-Standard gefunden werden [19].

Leider erfordert diese Methode eine rechenintensive FIR-Filterung bei hoher Samplingrate. Um die Komplexität des Modells gering zu halten, wird ein einfacherer Zusammenhang mit nur einer Zeitkonstante gesucht. Dazu ist in der Arbeit von Lin, Ambikairajah und Holmes [13] ein Ansatz zu finden, in welcher die Vormaskierung als entscheidungsgesteuerte akausale, bzw zeitgespiegelte IIR-Filterung 1. Ordnung durchgeführt wird (Gleichung 48, ähnlich wie Abbildung aber akausal). Um diese Filterung kausal zu ermöglichen müsste wieder auf ein FIR-Filter zurückgegriffen werden, mit entsprechend hoher Gruppenlaufzeit und Länge.

Die gewünschte Einfachheit für eine Echtzeit-Lösung kann nur in einem kausalen IIR-Filter 1. Ordnung mit einer Zeitkonstante von etwa 2 - 3 ms gefunden werden. Das entspricht zwar nur

sehr grob einem cos²-förmigen FIR-Filter, oder einem akausalen IIR Ausschaltsprung, kann aber durch entsprechende zeitliche Verzögerung des zugehörigen Bandpassignals die leichte Antizipation der Vormaskierung und die zeitliche Verschmierung eines Schallereignisses nachbilden. Ein normalisierter rekursiver Digitalfilter 1. Ordnung (*autoregressive first order section*) hat die Form:

$$H_{ARFOS}(z)=\frac{1-r_{ARFOS}}{1-r_{ARFOS}z^{-1}}. \tag{48}$$

Mit der Impulsinvarianztechnik ergibt sich der Radius r_{ARFOS} aus:

$$r_{ARFOS}=e^{-\frac{1}{\tau\cdot f_s}}. \tag{49}$$

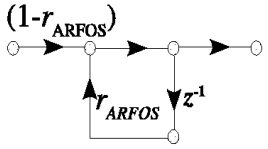

Abbildung 30: autoregressiver Filter 1. Ordnung

Diese Struktur benötigt 2 Multiplikationen, 1 Speicher und 2 Koeffizienten.

Die Implementierung der Vormaskierung könnte auch gänzlich weggelassen werden, es wird aber empfohlen sie durch geeignete Wahl der Zeitkonstante in der Mittelung der Pegelbestimmung mit Nichtlinearität zu integrieren. Werden phasendrehende Methoden zur Pegelbestimmung verwendet, so wird darauf hingewiesen, dass die Amplitudeneinhüllenden von Gammatone-Impulsantworten bei tiefen Mittenfrequenzen bereits einen Verlauf besitzen, welcher der Vormaskierung gut entspricht. Dies gilt bei der Verwendung von 1 Bark breiten Gammatone-Filtern für Mittenfrequenzen unter etwa 1 kHz, bzw. unter 10 Bark).

Werden statt Amplitudengrößen Leistungsgrößen verwendet, so muss die Zeitkonstante der Mittelung halbiert werden, um qualitativ dasselbe Ergebnis zu erhalten.

3.5.2 Modell für die Nachmaskierung

Die Zeitkonstante der Nachmaskierung ist abhängig vom betrachteten Frequenzband, dem darin vorhandenen Signalpegel, der Signalform und der Signaldauer, außerdem kann sie aufgrund großer Zeitkonstanten nichtmehr als lineare Filterung dargestellt werden, diese würde wahrnehmbare

Signalspitzen zu stark verschleifen und eine große Gruppenlaufzeit mit sich bringen. Aus der Arbeit von Frank Baumgarte [20] (Abbildung 31) kann man eine Struktur zur entscheidungsgesteuerten Mittelung entnehmen, die mit zwei Zeitkonstanten arbeitet.

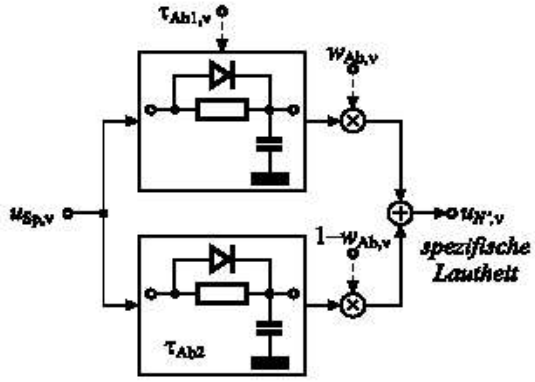

Abbildung 31: Struktur zur Nachmaskierung aus zwei Teilen, Frank Baumgarte [20].

Diese Struktur ist bereits relativ effizient, trotzdem greifen wir hier auf ein einfacheres Modell aus der Arbeit von Lin, Ambikairajah und Holmes zurück [13], welches die Nachmaskierung als entscheidungsgesteuerte IIR-Filterung 1. Ordnung mit nur einer Zeitkonstante beschreibt (Gleichung 49), also nur einem Teilfilter von oben, siehe Abbildung 30.

$$a^2_{k,\,pre}[n] = max\left\{a^2_k[n],\ \ (1-r_{pre})\cdot a^2_k[n] + r_{pre}\cdot a^2_{k,\,pre}[n+1]\right\}. \tag{50}$$

$$a^2_{k,\,pre+post}[n] = max\left\{a^2_{k,\,pre}[n],\ \ (1-r_{post})\cdot a^2_{k,\,pre}[n] + r_{post}\cdot a^2_{k,\,pre+post}[n-1]\right\}. \tag{51}$$

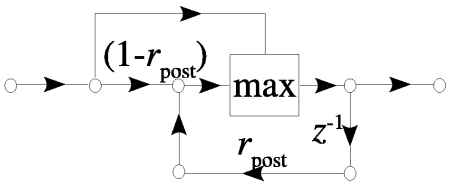

Abbildung 32: autoregressiver Filter 1. Ordnung mit Mittelung fallender Flanken und Durchlass steigender Flanken

Die nicht kausale Vorverdeckung nach Gleichung 48 ist nicht für *realtime* Implementationen geeignet, die Nachverdeckung kann mit gutem Erfolg eingesetzt werden. Die Zeitkonstanten der Nachmaskierung sind bei 1 Bark etwa $\tau_1 = 40$ ms und bei 20 Bark etwa $\tau_{20} = 4$ ms. Eine Formel zur

Berechnung der Zeitkonstanten in den einzelnen Frequenzbändern wurde aus einer Geradengleichung für 1/*f* gebildet und lautet (PEAQ [19]):

$$\frac{\tau_{post}}{[ms]} = \tau_{20} + \frac{Tfm_{Bark}^{-1}\{1\}}{f}(\tau_{20} - \tau_1). \tag{52}$$

Eine Formel zur Berechnung der Zeitkonstanten in den einzelnen Frequenzbändern, welche längere Mittelungskonstanten verwendet wurde aus einer Geradengleichung für 1/τ gebildet und lautet:

$$\frac{\tau_{post}}{[ms]} = \frac{1}{\frac{9}{32}\cdot\frac{f}{10000} + \frac{71}{3200}}. \tag{53}$$

Die zweite Gleichung beschreibt in der *Bark*-Skala beinahe einen linearen Abfall der Mittelungszeit von tiefen zu hohen Barkbändern, die erste Gleichung beschreibt im Vergleich dazu einen exponentiellen oder reziproken Abfall.

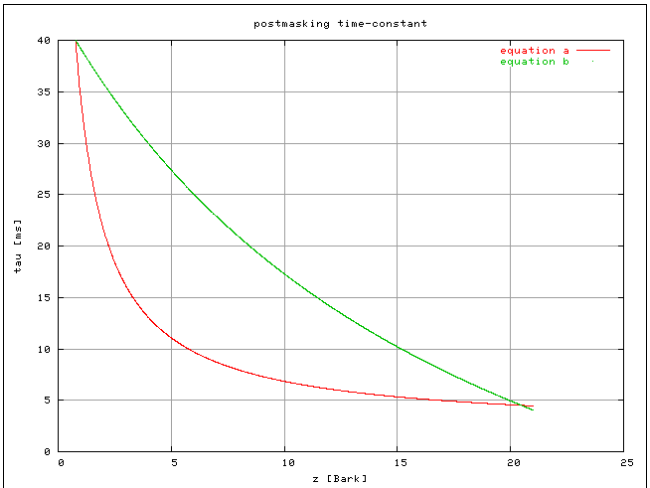

Abbildung 33: zwei Varianten für die Zeitkonstante zur Nachverdeckung

4 Digitale Signalverarbeitung in einer auditiven Domäne

Im vorigen Abschnitt wurde eine einfache auditive Signalanalyse beschrieben, die nun als Umgebung (*front-end*) für digitale Signalverarbeitung genützt werden kann. Dieser Abschnitt beschreibt im Wesentlichen einen Algorithmus zur Störgeräuschunterdrückung in dieser auditiven Umgebung. Zudem wird ein Algorithmus zur Unterdrückung von störendem Rückkopplungs-Pfeifen (*Howling*) beschrieben, der ebenfalls in dieser Domäne arbeitet.

4.1 Identifikation und Schätzung von Störsignalen (Leistungsdichtespektrum)

Um in einem Sprachsignal enthaltene Störungen unterdrücken zu können, ist es nötig diese zu identifizieren und zu quantifizieren. Zur Identifikation von Störsignal und Nutzsignal müssen Regeln aufgestellt werden, die sich auf bestimmte Charakteristika beider Signale beziehen. Auf ein Sprachsignal treffen folgende Eigenschaften zu:

- Schallpegel: Eine sprachliche Äußerung (*utterance*) hat meist einen Anfang und ein Ende, wir nehmen zudem an, dass in fließender Sprache nur zwischen solchen *utterances,* die aus mehreren fließend aneinander gereihten Wörtern bestehen, Pausen auftreten.
  - Der Beginn einer Äußerung müsste in den meisten Fällen durch eine rasche Anhebung der Signalleistung erkennbar sein. (vgl. Martin [45])
  - Sprachsegmente selbst zeichnen sich durch starke dynamische Amplitudenmodulationen aus, wobei der mittlere Pegel höher liegen sollte als jener der Störung. (vgl. Hess [46])
  - Sprachsegmente überschreiten niemals eine bestimmte Länge, es treten wiederholt Pausen auf.
- Das Schallfeld eines Sprachsignals hat, wenn das Mikrofon innerhalb des Hallradius vom Sprecher aufgestellt ist, kohärente Wellenfronten, die folgendes ermöglichen:
  - Definition eines erlaubten Aufenthaltsbereichs des Sprechers und Erkennung, ob ein Schallgeschehen jenen erlaubten Richtungen zugeordnet werden kann. (vgl. Zhang und Hansen [4])
  - Untersuchung der Kohärrenz des Schallfeldes. Bei Sprache wird große Kohärrenz erwartet, für Nachhall und Störgeräusche hingegen nicht. (vgl. Falch [2] und Bitwave [3])
- Methoden der *computational auditory scene analysis* (CASA) könnten hier auch ihre

Verwendung finden, benötigen aber viel Rechenaufwand. (Harmonizität, *double-vowel segregation*, *auditory grouping*, etc.)

- Sprachsignalspektren haben Tiefpasscharakteristik und weisen Formanten (spektrale Überhöhungen) auf, anhand welcher es möglich ist Sprachlaute zu erkennen (spektrale Gestalt).

Störgeräusche sind ausgezeichnet durch:

- Schallpegel: Meist besitzen Störgeräusche langsam veränderliche spektrale Signalleistungen. Man kann also überall dort, wo das Nutzsignal keine Spektralkomponenten besitzt und inaktiv ist, durch Mittelung die spektrale Gestalt (Periodogramm) der Störung ermitteln. Durch Halten gefundener spektraler Störleistungswerte kann die spektrale Gestalt während aktiven Spektralkomponenten des Nutzsignals geschätzt werden (vgl. Doblinger [47]).
- Rauschähnliche Störsignale besitzen keine harmonische Struktur und haben eher flache Spektren, *howling* oder Einstreuungen können jedoch auch schmalbandiger sein als Sprache. Durch ein Maß der spektralen Flachheit (SFM, *spectral flatness measure*) kann eine Unterscheidung getroffen werden. (vgl. Thiemann [24])
- Transiente Störungen können dann identifiziert werden, wenn sie nur sehr kurz auftreten und eine sehr hohe Leistung besitzen. Es könnte etwa ein Schwellwert bezogen auf die mittlere Sprachsignalleistung (vgl. *„bump noise*“ in Zhang und Hansen [4]) und eine Zeitschranke als Kriterium für laute transiente Schallereignisse verwendet werden.
- Hintergrundgeräusche im Auto haben stark ausgeprägten tieffrequenten Inhalt im Bereich <200 Hz,. Die höchsten enthaltenen Frequenzen gehen etwa bis 2,3 kHz. (vgl. [48])
- Störsignale können im Raum entweder diffuse oder gerichtete Wellenausbreitung haben. großflächige Körperschall-Störsignalquellen oder Nachhall besitzen meistens diffuse Schallfelder, störende Sprach- oder sonstige Quellen bilden oft gerichtete Schallfelder aus, die durch *beamforming* räumlich gefiltert werden können. (vgl. Zhang und Hansen [4])

Werden geeignete Methoden gefunden, um Störsignale zu identifizieren, so kann das Störsignalspektrum durch Mittelung der spektralen Leistungen (Periodogramm) während der Störsignalaktivität gefunden werden (vgl. Cohen [32]).

In folgenden Absätzen werden die Mittelung der spektralen Störsignalleistungen und einige

Varianten zur Steuerung der Mittelungsparameter erklärt.

4.1.1 Mittelung der Störsignalleistung zur Schätzung des Leistungsdichtespektrums

Um das Leistungsdichtespektrum eines Signals $a\,[n]$ zu berechnen, muss über jedes spektrale Amplitudenquadrat $a_k^2[m]$ des Frequenzbandes k ein Erwartungswert $E\{a_k^2[m]\}$ gebildet werden. Unter Annahme eines stationären Signals kann dieser Erwartungswert durch einen Zeitmittelwert über einen möglichst großen Zeitraum berechnet werden. Geht man von einem instationären Signal aus, das kurzzeitig stationäre Spektralamplituden besitzt, kann der Erwartungswert durch eine rekursive zeitliche Mittelung genähert berechnet werden. Dabei ergibt sich ein Periodogramm, eine Schätzung des Leistungsdichtespektrums (vgl. Cohen [32]). Die Zeitkonstante der Mittelung sollte in dem Bereich liegen, in welchem das Signal ein stationäres Spektrum besitzt. Zudem wird hier vorgeschlagen, die Zeitkonstante zur Mittelung als Vielfaches der Zeitkonstante der Nachverdeckung des menschlichen Gehörs zu wählen. Damit entspricht die Schätzung des Störsignalspektrums besser der subjektiv wahrgenommenen Störung. Passende Zeitkonstanten, welche in etwa der Stationarität eines Störgeräusches entsprechen liegen etwa bei 100-200 ms (vgl. Meyer *et al* [49]). Eine einfache rekursive Mittelung $E\{a_k^2[m]\}$ der Spektralkomponente $a_k^2[m]$ in diesem Sinne ist:

$$E\left\{a_k^2[m]\right\} \;\simeq\; r_{k,avg}\cdot E\left\{a_k^2[m-1]\right\}+(1-r_{k,avg})\cdot a_k^2[m]. \tag{54}$$

Dabei kann die Mittelungskonstante $r_{k,avg}$ wieder in Form einer Zeitkonstante angegeben werden (Impulsinvarianztechnik):

$$r_{k,avg}=e^{-\frac{1}{f_s\cdot\tau_{k,avg}}}. \tag{55}$$

Die Gleichung 54 kann folgende Operationen in Abhängigkeit von $r_{k,avg}$ durchführen:

1. rekursives *Mitteln* von $a_k^2[m]$, wenn $0 < r_{k,avg} < 1$, $(\tau_{k,avg} > 0)$
2. *Halten* des letzten Ausgangswertes, wenn $r_{k,avg} = 1$, $(\tau_{k,avg} = \infty)$
3. *Durchlassen* des Signals, wenn $r_{k,avg} = 0$, $(\tau_{k,avg} = 0)$
4. *Vergrößern* (Skalieren) des letzten Ausgangswertes, wenn $r_{k,avg} < 1$, $(\tau_{k,avg} < 0)$ und $a_k^2[m] = 0$

Zur Schätzung der spektralen Störsignalleistung $E_n\{a_k^2[m]\}$ muss der Mittelungsparameter $r_{k,avg}$ so eingestellt werden, dass zum Störsignal gehörende spektrale Leistungen rekursiv gemittelt werden. Während Sprecheraktivität soll der ermittelte Schätzwert hingegen konstant gehalten werden.

Um eine korrekte Steuerung von $r_{k,avg}$ zu ermöglichen, gibt es prinzipiell zwei Mechanismen, welche die Mittelung der spektralen Störsignalleistungen zwischen *Halten* und *Mitteln* (und ggf. *Vergrößern*, siehe Arslan *et al* [50]) umschalten.

1. Dazu kann eine Sprach- und Sprachpausenerkennung VAD (*voice activity detection: histogram method, level statistics*) verwendet werden. Diese VAD kann auf ein Leistungsmaß des Gesamtsignals basierend bestimmt werden. Liegt eine energiearme Sprachpause vor, wird das Störsignalspektrum in allen Frequenzbändern durch *Mitteln* geschätzt, in Sprachsegmenten werden alle ermittelten spektralen Störsignalleistungen durch *Halten* bestimmt.

   Der Vorteil dieser Methode liegt in der Robustheit während Sprachpausen, da schmalbandige, energiearme Veränderungen nicht zu Detektion von Sprache führen. Bei sehr lauten Umgebungsgeräuschen ermöglicht dieses Verfahren allerdings keine genaue Störsignalschätzung, da niederpegelige Sprache oder energiearme Sprachlaute nicht immer richtig detektiert werden können.
2. Eine Verfeinerung der Störsignalschätzung ergibt sich, wenn jedes Frequenzband individuell auf Sprachinhalt geprüft wird, weil dadurch eine häufigere Aktualisierung der Schätzung möglich wird (*minimum statistics, harmonic tunneling, quantile-based, histogram-method, minima controlled recursive averaging*). Ob ein Frequenzband Sprachinhalt besitzt ergibt sich aus der Untersuchung der Spektralamplitude. Werden lokal Signalamplituden vorgefunden, die einem Störsignal zugeordet werden können (klein, häufig, länger als Sprachsegmente, ...), so wird die lokale Schätzung der Störsignalamplitude durch *Mitteln* getroffen. Wird Sprachinhalt erkannt, ergibt sich der ermittelte Schätzwert wieder durch *Halten*.

   Dieses Verfahren arbeitet auch in schwierigen Verhältnissen mit großen Störgeräuschpegeln noch so, dass ein Großteil der spektralen Sprachinhalte noch gut identifiziert werden kann. In Sprachpausen können allerdings schmalbandige Störsignalspitzen als lokaler Sprachinhalt fehldetektiert werden, womit hier *musical noise* vermehrt auftreten kann.

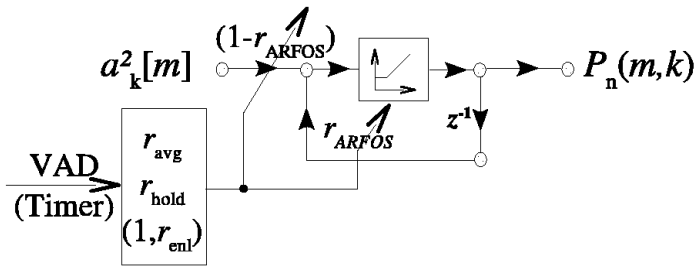

Abbildung 34: rekursive Mittelung der spektralen Störleistung $P_n(m,k)$ zur Zeit m im Frequenzband k

In sehr schlechten Verhältnissen ist eine lokale Detektion von Sprachinhalt vorzuziehen, die kleinen Signaländerungen ermöglicht noch als Sprachinhalt erkannt zu werden. Bei sehr guten Verhältnissen ist eher die VAD vorzuziehen, die in den Sprachpausen wenigere Fehldetektionen aufweist.

Berücksichtigung der zeitlichen Maskierung: Werden die zeitlich maskierten spektralen Leistungen $a_{\text{k-pre+post}}^2[m]$ verwendet, um die Störsignalleistung zu schätzen, ergibt sich bei impulsiven Störungen ein höherer Störsignalpegel. Dies hat in weiterer Folge positive Auswirkungen auf die spektrale Subtraktion, indem impulshaltige Störsignalpegel mit erhöhter Leistung geschätzt werden. Die daraus resultierende Übersubtraktion kann *musical noise* verringern (siehe Übersubtraktion in Berouti *et al* [51]).

Die Schätzung der spektralen Störsignalleistung benötigt bei N-kanaliger Spektralanalyse N Additionen, 2N Multiplikationen, N Koeffizienten und N Speicher. Soll auch eine rekursive Vergrößerung vorgesehen werden, so ergeben sich 2N Koeffizienten.

4.1.2 Sprach- und Sprachpausendetektion (VAD)

Eine globale Sprache-Pause-Detektion kann über Betrachtungen der Gesamtsignalleistung (oder Signalpegel) und dessen zeitliche Änderung durchgeführt werden. Dazu gibt es einige Methoden:

- *Histogramm-Methode* nach Hess [46] (vgl. Vary/Heute/Hess [38]):

  Bildet man ein Histogramm der Gesamtsignalleistung, das sich zeitlich mit dem Signal verändern kann (z.B. indem man ein gleitendes Analyseintervall untersucht), kann gezeigt werden, dass der Signalpegel mit der größten Auftrittshäufigkeit zum Störsignal gehört. Dieses Maximum bildet sich deshalb aus, weil Sprache im Gegensatz zum Störgeräuschpegel meist stark zeitveränderliche Pegelmodulationen aufweist, und andererseits die Sprachpausen meist mehr Zeit in Anspruch nehmen als die Sprache selbst (vgl. Vary/Heute/Hess [38]). Nun kann aus dem gefundenen

Maximum ein Schwellwert berechnet werden. Aus dem Vergleich des aktuellen Pegels mit diesem Schwellwert ergibt sich die Sprache/Pause-Unterscheidung. Ein Problem dieser Statistik ist allerdings, dass energiearme Sprachlaute wie Frikative oft nicht erkannt werden. Eine genauere Entscheidung erhält man, wenn man auch für die Signalpegel des differenzierten Signals ein Histogramm bildet, und den daraus gebildeten Schwellwert zum Vergleich mit dem aktuellen Signalpegel des differenzierten Signals heranzieht.

- VAD über Schätzung der Gesamtstörsignalleistung mit adaptiven Schwellwerten für Sprache und Geräusch (vgl. Bitwave Patent [3], Doblinger [47]):

  Dabei werden 2 Schwellwerte für die Gesamtsignalleistung gebildet, die adaptiv mit der Schätzung einer gesamten Störsignalleistung ($P_n[m]$ *noise floor*) verändert werden. Die Unterschreitung des ersten Schwellwertes $T_n[m]$ für Störgeräusche gibt an, dass der aktuelle Gesamtsignalpegel $P_x[m]$ einem Störsignal zugeordnet wird, der zweite, höhrergelegene Schwellwert für Sprachaktivität $T_s[m]$ gibt bei Überschreitung an, dass Sprachaktivität vorliegt.

  Bei Störsignalaktivität wird durch *Mitteln* (vgl. Gleichung 54) der Gesamtsignalleistung $P_x[m]$ eine gesamte Störsignalleistung $P_n[m]$ bestimmt. Liegt Sprachsignalaktivität vor, wird die Gesamtstörsignalleistung $P_n[m]$ durch *Halten* geschätzt. Dauert die Sprachaktivität länger als ein festgelegtes maximales Intervall von etwa T_{cc} = 2-7 s (vgl. Bitwave [3]) an, so wird die geschätzte Störsignalleistung $P_n[m]$ so lange langsam durch *Vergrößern* erhöht, bis wieder eine Sprachpause detektiert wird. Dadurch kann auch einem rasch steigenden Störsignalpegel noch langsam gefolgt werden.

Der Algorithmus zur VAD durch Schätzung einer Gesamtsignalleistung mit adaptiven Schwellwerten besitzt den geringsten Berechnungsaufwand in der Implementierung. Als Gesamtsignalleistung kann entweder der Teager-Energy Operator, der gegenüber Störgeräuschen unempfindlich ist [48], oder die Lautheit verwendet werden.

In den folgenden Abschnitten werden diese beiden Leistungsgrößen, Teager-Energy Operator und Lautheit, kurz beschrieben.

4.1.2.1 Teager-Energy Operator (TEO)

Ein mögliches Maß für die Signalleistung des gesamten gestörten Sprachsignals $P_x[n]$ ist der nichtlineare Teager-Energy Operator (TEO) mit der Form $T(x)=x^2[n]-x[n-1]x[n+1]$. Wird der

Ausdruck T(x) rekursiv gemittelt kann man damit den Erwartungswert schätzen [48]:

$$P_x[n]=E\{T(x[n])\} \;=\; r_{xx}[0]-r_{xx}[2] \;\simeq\; r_{n,avg}\cdot E\{rT(x[n-1])\}+(1-r_{n,avg})\cdot T(x[n]) \qquad (56)$$

Dabei sieht man, dass tieffrequente Signalanteile, mit einer starken Korrelation über 3 Samples (*Anm.: gleiches Vorzeichen in der Korrelation!*), aus dem berechneten Leistungswert abgezogen werden. In [48] wird gezeigt, dass Sprachsignale in diesem Energiemaß stärker bewertet werden als Störgeräusche im Auto. Das liegt daran, dass Fahrzeuggeräusche im Gegensatz zur Sprache sehr großen tieffrequenzen Inhalt besitzen. Der Teager-Energy-Operator bewertet den tieffrequenten Energieanteil des Signals nur wenig, der hochfrequente Energieanteil wird normal bewertet.

Eine zusätzliche Diskriminierung von Störgeräuschen kann erzielt werden, wenn der TEO in Frequenzbändern einer Spektralanalyse berechnet wird und durch Summation der Absolutbeträge zusammengefasst wird [48].

In der Praxis sollte der ermittelte Leistungswert nach der rekursiven Mittelung (vgl. Gleichung 54) unterabgetastet werden mit dem Zeipunkt $m = \mathrm{L}\cdot n$ mit dem Unterabtastungsfaktor L. Eine passende Zeitkonstante für die Mittelung wäre etwa 50 - 70 ms.

Die Berechnung des gemittelten Teager-Energy Operators benötigt 3 Speicherstellen, 4 Koeffizienten, 2 Additionen und 4 Multiplikationen.

4.1.2.2 Lautheit

Als Maß für die gesamte Leistung des gestörten Sprachsignals kann auch der Lautstärkeeindruck herangezogen werden, welcher der auditiven Signalanalyse entstammt und sowohl die simultane und zeitliche Verdeckung beinhaltet. Diese Lautheit würde in diesem Fall zum Zeitpunkt *m* über die Additionen der Leistungen (unter Berücksichtigung der zeitlichen Maskierung) in den Einzelkanälen der auditiven Filterbank berechnet werden, siehe auch Pflüger [6] Zwicker [8] (vgl. Gleichung -Zeitmaskierung, hier mit *Downsampling* $m = \mathrm{L}\cdot n$).

$$P_x[m]=\sum_{k=1}^{N}\left(a^2_{k,\,pre+post}[m]\right)^{0,3.} \qquad (57)$$

Ein Vorteil dieses Maßes gegenüber dem Teager-Operator ist, dass es besser die menschliche Lautheitsempfindung nachbildet. Das menschliche Gehör ist besonders empfindlich auf Sprache, weshalb angenommen werden darf, dass auch dieses Maß ausgezeichnete Eigenschaften in der Sprachdetektion besitzt (vgl. [9]). Die Außen-Mittelohr-Übertragungsfunktion zum Beispiel blendet bereits jene tieffrequenten Störgeräusche aus, welche im Fahrzeug so dominant sind. Allgemein gilt

bei Verwendung des Lautheitsmaß, dass Sprachsignale dann detektiert werden, wenn die Lautheit größer ist als jene bei reinen Störungen.

Die Berechnung der Lautheit aus den Amplitudenquadraten $a_{k,\,pre+post}$ benötigt in einer M-Kanaligen auditiven Filterbank M-1 Additionen und M Potenzfunktionen.

4.1.2.3 VAD über Schätzung einer gesamten Störsignalleistung mit adaptiven Schwellwerten(noisefloor)

Eine Erkennung der Sprachaktivität (VAD, *voice activity* detection) zum Zeitpunkt m kann über eine Untersuchung der gesamten Signalleistung erfolgen. Dazu werden adaptiv Schwellwerte für die Signalleistung gebildet, anhand welcher sich die Entscheidung ergibt. Dies ist die einfachste Variante zur Erkennung der Sprecheraktivität. Datz werden zwei Schwellwerte $T_n[m]$ und $T_s[m]$ und eine Schätzung der gesamten Störsignalleistung $P_n[m]$ aus der Signalleistung $P_x[m]$ zum Zeitpunkt m gewonnen. In folgenden Punkten wird das Verhalten des Verfahrens (vgl. Bitwave Patent [3]) beschrieben. Dabei werden folgende Annahmen verwendet:

- Eine Sprachpause liegt vor, wenn die momentane Leistung $P_x[m]$ unter dem Schwellwert $T_n[m]$ für Störsignale liegt. In dieser Zeit kann die Störsignalleistung $P_n[m]$ durch Mittelung geschätzt werden. Aus dem Schätzwert $P_n[m\text{-}1]$ ergeben sich die neuen Schwellwerte $T_n[m]$ und $T_s[m]$ für Stör- und Nutzsignal durch jeweilige Skalierung im Bereich 1.05 bis 2.
- Eine Spektralkomponente mit Sprachinhalt liegt vor, wenn die momentane Leistung $P_x[m]$ über dem Schwellwert $T_s[m]$ für Sprache liegt. In diesem Fall kann die Störsignalleistung des vorangegangenen Analysezeitpunktes als aktueller Schätzwert $P_n[m] = P_n[m\text{-}1]$ verwendet werden.
- Eine sprachliche Äußerung (*utterance*) dauert etwa T_{cc} = 2-7 s. Liegt die momentane Leistung länger als diese Zeit überhalb des Schwellwertes $T_s[m]$ für Sprache, so muss davon ausgegangen werden, dass sich Störsignalleistung erhöht hat und eine Sprachpause vorhanden ist. Das aktuelle Schwellwertepaar $T_n[m]$ $T_s[m]$ und die Störsignalleistung $P_n[m]$ sollten dann über eine Skalierung des vorangegangenen Schätzwertes $P_n[m] = r_{enl}\, P_n[m\text{-}1]$ erhöht werden.
- Die Mittelung des Leistungsdichtespektrums des Störsignals in den Frequenzbändern kann über diese VAD gesteuert werden. Dazu können dem Gehör angepasste Zeitkonstanten verwendet werden. Es bietet sich an Vielfache der Zeitkonstante der Nachmaskierung einzusetzen. Bänder tiefer Frequenzen sollten in etwa eine Zeitkonstante von 200 - 500 ms verwenden.

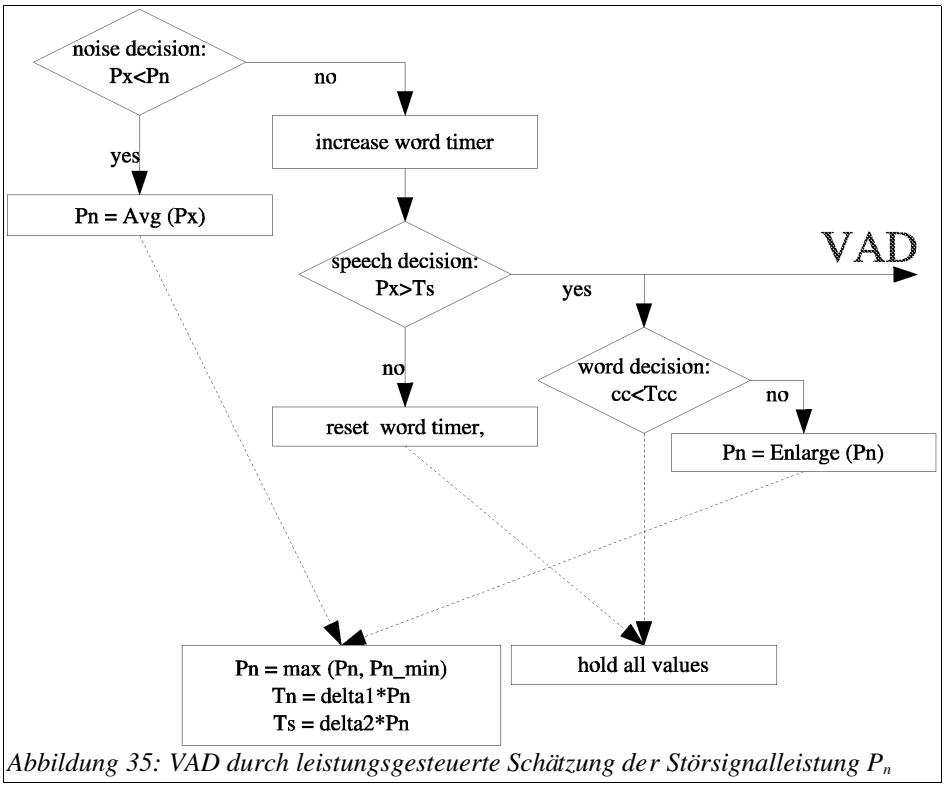

Abbildung 35: VAD durch leistungsgesteuerte Schätzung der Störsignalleistung P_n

Das Verfahren benötigt 4 Speicherstellen, 6 Koeffizienten, 1-2 Additionen und 4 Multiplikationen. Es sind etwa 5 Vergleichsoperationen durchzuführen.

$$r_{ARFOS} = \begin{cases} r_{avg} & wenn\, VAD = 0 \\ 1 & wenn\, VAD = 1. \end{cases} \tag{58}$$

4.1.3 Steuerung der Störsignalschätzung durch frequenzselektive Detektion des Sprachinhalts (ohne explizite VAD)

Werden die Signalleistungen aller Analysefrequenzbänder separat untersucht, kann in jedem Frequenzband eine Erkennung der Signalaktivität (Sprachinhalt) durchgeführt werden. Auf diese Weise kann die Störsignalleistung lokal geschätzt werden. Der Schätzwert lässt auch während eines Sprachsegmentes in jenen Teilen des Spektrums eine Aktualisierung zu, die keinen Sprachinhalt führen. Der Nachteil dabei ist, dass einzelne Frequenzbänder in Sprachpausen falsche Detektionen von Sprachinhalt aufweisen können. Es gibt folgende Methoden frequenzselektiv den Sprachinhalt zu detektieren:

- *Histogramm-Methode* nach Hirsch und Ehrlicher [52] (*modal*):

Die Histogramm-Methode kann lokal auf die individuellen Leistungen der Einzelfrequenzbänder angewendet werden. Dabei wird für jedes Analysefrequenzband ein Histogramm erstellt, dessen häufigste Leistung (*Modalwert*) zur Bildung eines Schwellwerts für Sprachinhalt verwendet wird. Das Problem dabei ist allerdings, dass eine ausreichend feine Quantisierung der Leistungen gefunden werden muss. Diese Methode funktioniert am besten, ist allerdings die aufwändigste und jene, die am meisten Speicher benötigt.

- Schätzung der spektralen Störsignalleistungen mit frequenzselektiven adaptiven Schwellwerten für Sprache und Geräusch (vgl. Bitwave-Patent [3], *Minimum Statistik* von Doblinger [47], Mittelung in Arslan *et al* [50], *mittel*):

  Dazu werden für jedes Frequenzband k aus der individuell geschätzten Störsignalleistung zwei Schwellwerte gebildet. Der erste über der lokalen Störsignalleistung $P_{\mathrm{n}}(m, k)$ liegende Schwellwert für Störsignale $T_{\mathrm{n}}(m, k)$ indiziert bei Unterschreitung durch die Signalleistung $P_{\mathrm{x}}(m, k)$ Störsignalinhalt im Frequenzband k und schaltet die Schätzung der Störsignalleistung auf „Mitteln“. Der zweite über dem Schwellwert für Störsignale liegende Schwellwert $T_{\mathrm{s}}(m, k)$ für Sprache indiziert bei Überschreitung durch die Signalleistung $P_{\mathrm{x}}(m, k)$ Sprachinhalt im betrachteten Frequenzband und schaltet die Schätzung der Störsignalleistung solange auf „Halten“, bis die maximale Zeit T_{cc} für Sprachinhalt überschritten ist. Liegt die Signalleistung $P_{\mathrm{x}}(m, k)$ des Frequenzbandes k länger als erlaubt über dem Schwellwert $T_{\mathrm{s}}(m, k)$ für Sprache, wird die Schätzung der Störsignalleistung auf „Vergrößern“ umgeschaltet, bis der Schwellwert $T_{\mathrm{s}}(m, k)$ das erste Mal unterschritten wird.

  Die Struktur dieser einzelnen Steuerungen sieht ähnlich aus wie jene der VAD. Die ermittelte Störsignalleistung $P_{\mathrm{n}}(m, k)$ kann hier, im Gegensatz zum *noisefloor* $P_{\mathrm{n}}[m]$ der VAD, direkt als spektrale Störsignalschätzung weiterverwendet werden. Dadurch ergeben sich ein paar Vereinfachungen im Schaltbild.

- *Minimum-Statistik* von Martin [45]: Martin verwendet eine rekursive Mittelung der Leistungsgrößen (*optimal smoothing*), deren Mittelungskonstante in Abhängigkeit vom gemittelten *a posteriori* SNR und einigen Korrekturtermen bestimmt wird. Die gemittelte Leistungsgröße folgt nun starken Änderungen bei großen Leistungen schnell (Sprache), Änderungen im Bereich der Leistungsminima werden stark gemittelt und unterdrücken damit ausreißende Werte für die darauffolgende Minumumsuche. Die Suche nach dem Minumum

(*minimum statistics*) innerhalb eines gleitenden Analysefensters ergibt dann die spektrale Störsignalleistung. Dieses aufwändigere Verfahren kann auch Änderungen im Störsignalpegel rasch folgen.

- *Quantile-basierte* Störsignalschätzung nach Stahl, Fischer und Bippus (*median/quantil*) [53]: Amplitudenwerte im Analysezeitraum, der etwa der Länge einer *utterance* besitzt, werden nach aufsteigender Reihenfolge sortiert. Es wird davon ausgegangen, dass - unabhängig von der Sprachaktivität - ein Frequenzband nur einen bestimmten Anteil q an Sprachinhalt zu Störinhalt besitzt. Wählt man jenen Leistungswert aus, der in der sortierten Reihe mit M Einträgen an jener Stelle sitzt (*q-Quantil* / $q = 0{,}5$: *Median*), die diesen Anteil, also $M \cdot q$ repräsentiert, so findet man einen Schwellwert zur Unterscheidung von Sprachinhalt von Störinhalt, wählt man $q = 0.5$, erhält man den Medianwert des Analysefensters als Schwellwert. Dieses Verfahren ist sehr aufwändig und benötigt viel Speicher.
- *Harmonic-Tunnelling*: In der Arbeit [54] von Ealey, Kelleher und Pearce werden mithilfe einer Detektion der Grundfrequenzperiode die Täler (*Tunnel*) zwischen den Harmonischen des Sprachspektrums gesucht. Innerhalb dieser Tunnel, also zwischen den Harmonischen der Sprache, kann nun auch während Sprachsegmenten die spektralen Leistungsdichte des Störsignals geschätzt werden. Dazu werden die Amplituden dieser Tunnel verfolgt. Bleibt Änderung bei der Amplitudenfortschreitung in einem kleinen Bereich, welcher einem Störsignal zugeordnet werden kann, werden die Amplituden gefundener Tunnel zur Störsignalschätzung verwendet. Dies erlaubt auch stark instationäre Hintergeräusche gut zu schätzen, dazu ist allerdings eine große Frequenzauflösung nötig, die bei der Analyse in kritischen Bandbreiten nicht gegeben ist.
- *Zeit-Frequenz Quantile-basiert*e Störsignalschätzung nach Evans und Mason [55]: Wie in dem originalen Artikel von Stahl *et al* werden hier Quantile-basierte Schätzwerte im Analysefenster für die spektralen Leistungsdichtewerte herangezogen. Hier werden zudem aber Quantil-basierte Schätzungen aus benachbarten Frequenzbänder mit dem aktuellen Leistungsdichtewerte kombiniert. Das bewirkt eine gewisse Mittelung über das Spektrum und nützt auch die Möglichkeiten des *Harmonic Tunnelling* aus. Bei der Analyse in Frequenzbänder mit kritischen Bandbreiten ist hier ebenfalls keine ausreichende Frequenzauflösung gegeben, um dieses Verfahren sinnvoll anwenden zu können.
- *Minima Controlled Recursive Averaging* - MCRA von Cohen [32]: Cohen verwendet in seiner

Arbeit ein rekursiv geglättetes Spektrum, in welchem spektrale Minima innerhalb eines gleitenden Zeitfensters gesucht werden. Anders als bei Martin werden gefundene Minima nicht direkt als Spektralschätzung verwendet. Das Verhältnis aus momentanem Spektralwert zu aktuellem Minimalwert wird mit einem Schwellwert geprüft, liegt es über der Schwelle, besitzt das betrachtete Frequenzband Sprachinhalt. Eine so gefundene Entscheidungsfunktion wird rekursiv gemittelt und als Wahrscheinlichkeit für Sprachinhalt im Frequenzband gesehen. Die so gefundene Wahrscheinlichkeit für Sprachpräsenz steuert dann schließlich den Mittelungsparameter einer rekursiven Störsignalschätzung, die direkt auf die Amplitudenquadrate des jeweils betrachteten Frequenzbandes angewendet wird.

4.1.3.1 Schätzung der spektralen Störsignalleistung mit frequenzselektiven adaptiven Schwellwerten

Eine sehr einfache Variante zur Auffindung von Schwellwerten $T_n(m,k)$ und $T_s(m,k)$ und Schätzung der spektralen Störsignalleistungen $P_n(m,k)$ aus den spektralen Signalleistungen $P_x(m,k) = a^2_k[m]$ wird in diesem Abschnitt beschrieben (vgl. Bitwave Patent [3]). Dabei werden folgende Annahmen verwendet:

- Eine Spektralkomponente ohne Sprachinhalt liegt vor, wenn die momentane Leistung $P_x(m,k)$ im Frequenzband k unter dem Schwellwert $T_n(m,k)$ für Störsignale liegt. In dieser Zeit kann die Störsignalleistung $P_n(m,k)$ durch Mittelung geschätzt werden. Aus dem Schätzwert $P_n(m\text{-}1,k)$ ergeben sich die neuen Schwellwerte $T_n(m,k)$ und $T_s(m,k)$ für Stör- und Nutzsignal durch jeweilige Skalierung im Bereich 1.05 - 2.
- Eine Spektralkomponente mit Sprachinhalt liegt vor, wenn die momentane Leistung $P_x(m,k)$ im Frequenzband k über dem Schwellwert $T_s(m,k)$ für Sprache liegt. In diesem Fall kann die Störsignalleistung des vorangegangenen Analysezeitpunktes als aktueller Schätzwert $P_n(m, k) = P_n(m\text{-}1,k)$ verwendet werden.
- Eine sprachliche Äußerung (*utterance*) dauert bei ganzen Sätzen etwa $T_{cc} = 2\text{-}7$ s. Liegt die momentane Leistung eines Frequenzbandes länger als diese Zeit überhalb des Schwellwertes $T_s(m,k)$ für Sprache, so muss davon ausgegangen werden, dass sich Störsignalleistung erhöht hat und kein Sprachinhalt vorhanden ist. Das aktuelle Schwellwertepaar $T_n(m,k)$ $T_s(m,k)$ und die Störsignalleistung $P_n(m,k)$ sollten dann über eine Skalierung des vorangegangenen Schätzwertes $P_n(m,k) = r_{enl} P_n(m\text{-}1,k)$ erhöht werden.

- Zur Mittelung der Störsignalleistung können dem Gehör angepasste Zeitkonstanten verwendet werden. Es bietet sich an Vielfache der Zeitkonstante der Nachmaskierung einzusetzen. Bänder tiefer Frequenzen sollten in etwa eine Zeitkonstante von 200 - 500 ms verwenden.

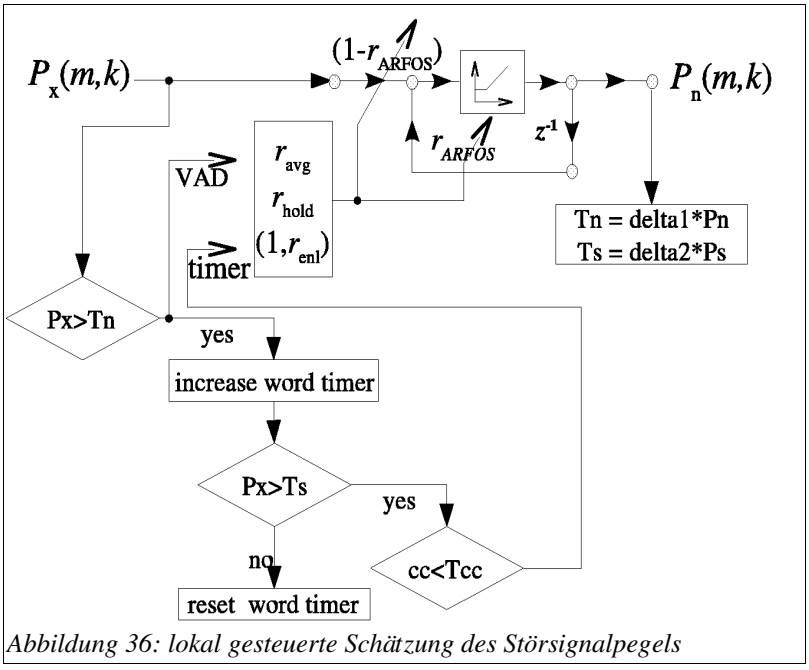

Abbildung 36: lokal gesteuerte Schätzung des Störsignalpegels

Das Verfahren benötigt in einer N-kanaligen Spektralanalyse etwa 4 N Multiplikationen, (1-2) N Additionen, 4 N Speicherstellen, 3 N+3 Koeffizienten und etwa 5 N Vergleichsoperatoren.

$$r_{k,ARFOS} = \begin{cases} r_{avg} & \left(P_x(m,k) < T_n(m,k)\right) \wedge \left(cc(m,k) < Tcc\right) \\ r_{hold} & \left(P_x(m,k) > T_n(m,k)\right) \wedge \left(cc(m,k) < Tcc\right) \\ r_{enl}, \; a_k^2[m] = 0 & \left(P_x(m,k) > T_n(m,k)\right) \wedge \left(cc(m,k) > Tcc\right). \end{cases} \tag{59}$$

4.2 Rauschunterdrückung

Rauschunterdrückung wird meist durch Subtraktion eines geschätzten Störspektrums vom Nutzsignalspektrum implementiert. Im Detail kann für gewöhnlich festgelegt werden, ob Betragsspektren oder Leistungsdichtespektren von einander abgezogen werden, oder variierte Subtraktionsregeln verwendet werden sollen (vgl. Abschnitte 4.2.2, 4.2.3, und [38]). Die Signalverarbeitung selbst stellt für das Nutzsignal nur eine spektrale Gewichtung mit

zeitveränderlichen Faktoren[5] $g(m,k)$ zum Analysezeitpunkt m im Frequenzband k dar, siehe Abbildung 37. Das Signal muss dazu einer Spektralanalyse unterzogen, gewichtet und wieder resynthetisiert werden (vgl. Vary/Heute/Hess [38]).

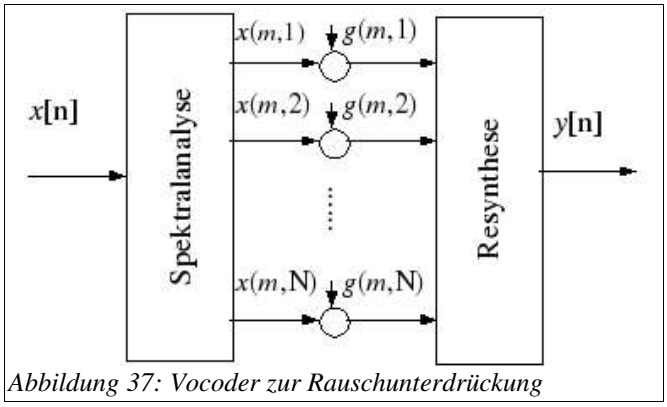

Abbildung 37: Vocoder zur Rauschunterdrückung

Nun ist dabei die Schätzung des Störsignalspektrums die erste Herausforderung. Lösungsansätze dazu wurden bereits im Abschnitt zuvor beschrieben.

Es stellt sich bei einer Implementierung einer spektralen Subtraktion mit den klassischen Methoden (vgl. Abschnitt 4.2.2 und 4.2.3) heraus, dass zwar der Störsignalpegel gesenkt werden kann, aber statt dessen neue Störungen, sog. *musical noise* oder *musical tones*, im Signal enthalten sind. Diese *musical tones* sind schmalbandige Störungen, die dann auftreten, wenn die Schätzung des Störsignals in einem Frequenzband kurzfristig vom tatsächlichen Störgeräusch überschritten wird. Die Verringerung von *musical tones* oder *musical noise* stellt die zweite Herausforderung bei der Sörgeräuschunterdrückung dar (vgl. Vary/Heute/Hess [38]).

Zum einen ist bei der Minderung von *musical noise* die Verwendung der Ephraim und Malah Subtraktionsregel (Abschnitt 4.2.4 bis 4.2.15) sehr hilfreich (vgl. Olivier Cappé [36]). Zum anderen kann auch eine Glättung der geschätzten Spektren dabei helfen, die entstehenden Artefakte breitbandiger werden zu lassen (vgl. Arslan *et al.* [50]). Außerdem zeigt sich, dass bei niedrigem Störanteil ein Überschätzen der Störsignalleistung (eine sog. Übersubtraktion) zur Reduktion von *musical noise* beiträgt (vgl. Berouti *et al.* [51]).

5 Die Wirkung der Spektralsubtraktion auf ein Frequenzband ist vergleichbar mit jener eines *noise gate*, das in der Audio-Signalverarbeitung auf verrauschte breitbandige Signale angewendet wird (*noise gate:* siehe dazu [43]). Energiearme Signalanteile werden dabei stumm geschaltet.

4.2.1 Verbesserte Rauschunterdrückung mithilfe auditiver Signalanalysen

Ein weiterer Ansatz, der zur Reduktion von *musical noise* führen kann, ist die Begrenzung der spektralen Gewichte auf die Mithörschwelle des menschlichen Gehörs (vgl. Nathalie Virag [56], Tsoukalas [39], Loizou [37], Thiemann [57][24]). Dies geschieht meist, indem parallel zum FFT-Spektrum die Mithörschwelle des ungestörten Nutzsignals iterativ bestimmt wird. Anhand dieser Mithörschwelle können dann die spektralen Gewichtsfaktoren zur Rauschunterdrückung berechnet werden. Diese Methode verhindert, dass unhörbare Störungen gänzlich unterdrückt werden, was zu unnötig starken Verzerrungen des Nutzsignals führen würde. Zudem ist in einigen Arbeiten ein Konzept zu finden, das von perfekter Störsignalunterdrückung abgeht. Dazu wird ein *noisefloor*-Parameter eingeführt, der zu starke Unterdrückung unterbindet. Damit werden Störsignale nur mehr bis zu einem gewissen Grad gedämpft und *musical noise* weiter unterdrückt (vgl. Gustaffson *et al* [58], Berouti *et al* [51], Virag [56]).

Übersubtraktion und *noisefloor*-Parameter können so kombiniert werden, dass zur Verringerung von *musical noise* eine starke Übersubtraktion und zur Begrenzung der Signalverluste ein großer *noisefloor*-Parameter eingesetzt wird.

Wird direkt eine auditive Filterbank zur Spektralzerlegung angewendet, erspart man sich die Berechnung der Mithörschwelle. Von Lin, Ambikairajah und Holmes existiert bereits eine Arbeit

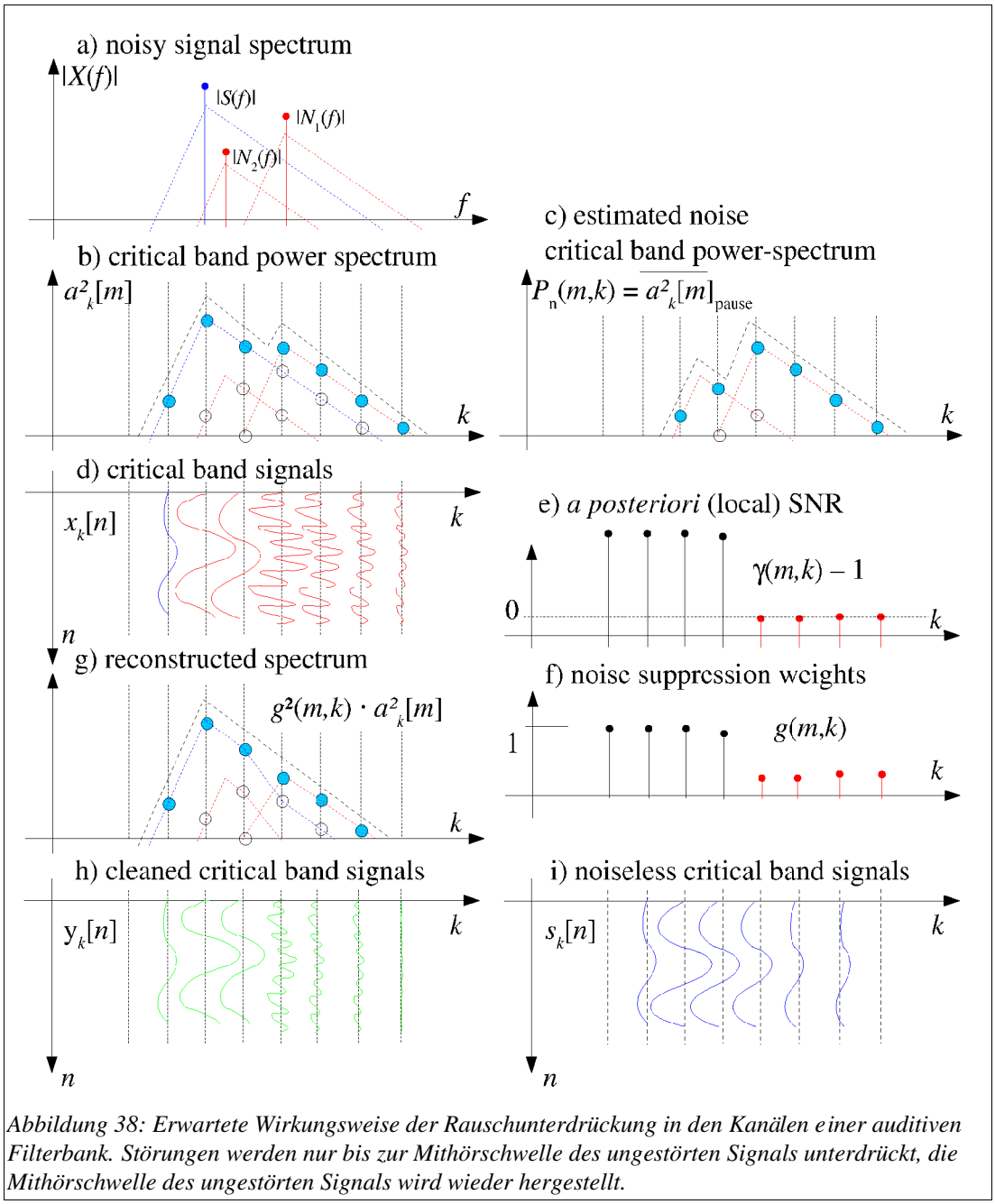

Abbildung 38: Erwartete Wirkungsweise der Rauschunterdrückung in den Kanälen einer auditiven Filterbank. Störungen werden nur bis zur Mithörschwelle des ungestörten Signals unterdrückt, die Mithörschwelle des ungestörten Signals wird wieder hergestellt.

mit diesem Ansatz und einer modifizierten Wiener-Subtraktionsregel [59].

Das Bitwave-Patent enthält im Gegensatz zur auditiven Filterung im Zeitbereich eine FFT-Analyse mit Bark-Transformation, siehe Abbildung 40. Mit einer auditiven Filterbank ergibt sich das

Strukturbild aus Abbildung 39.

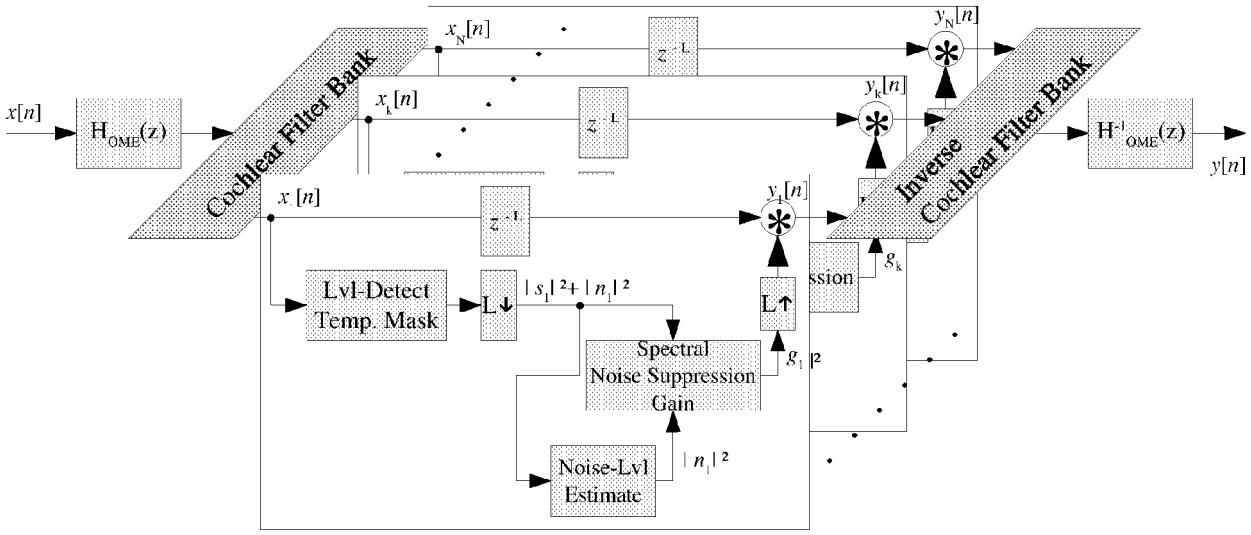

Abbildung 39: Rauschunterdrückung in einer auditiven Signalanalysestufe mit kochleärer Filterbank (z.B.: Gammatone-Filterbank)

In Abbildung 38 wird grafisch gezeigt, wie die Mithörschwelle des ungestörten Signals durch eine Rauschunterdrückung im auditiven Bereich wiederhergestellt werden kann.

Die Arbeit [23] von Irino erreicht mit einer auditiven Gammachirp-Filterbank eine Rauschunterdrückung, indem der Gammachirp-Asymmetrieparameter so gesteuert wird, dass Analyse- und Resynthesefilter zusammen eine vom SNR abhängige Dämpfung bewirken. Nach den Ergebnissen neuerer Arbeiten zeigt sich jedoch, dass die Gammachirp-Asymmetrie nicht lautstärkeabhängig ist (Irino und Patterson [44]). Auch aufgrund des erheblich hohen Rechenaufwands (zwei Analysebänke und eine Resynthesebank) wird dieser interessante Ansatz hier abgelehnt.

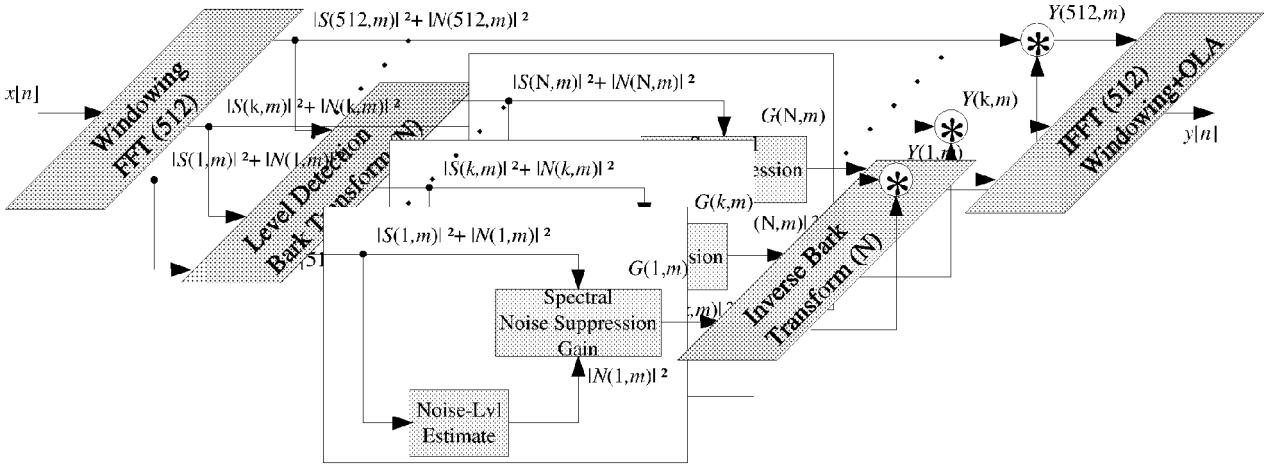

Abbildung 40: Rauschunterdrückung in einer Bark-Domäne mit FFT-Spektralzerlegung

4.2.2 Wiener-Filter

Zur Entstörung eines verrauschten Signals $x[n] = s[n] + n[n]$ mit dem Nutzsignal $s\,[n]$ und dem Störsignal $n\,[n]$ kann ein Wiener-Filter verwendet werden, siehe Abbildung 43.

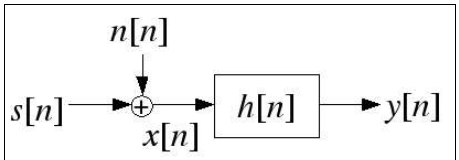

Abbildung 41: Wiener Filter ([38])

Der Ansatz für den Wienerfilter minimiert das erwartete Fehlerquadrat [38]:

$$E\left\{\left(y[n]-s[n]\right)^2\right\}=E\left\{\left(\sum_{n=-\infty}^{k=\infty} h[k]\cdot x[n-k]-s[n]\right)^2\right\}\rightarrow min. \tag{60}$$

Aus der Ableitung nach $h[n]$ und der Forderung, dass die Ableitung für den gesuchten Extremwert 0 werden muss, ergibt sich [38]:

$$\sum_{k=-\infty}^{k=\infty} h[k]\cdot\varphi_{xx}[n-k]=\varphi_{xs}[n]. \tag{61}$$

Hier sind mit φ_{xx} die Autokorrelation des gestörten Signals und mit φ_{xs} die Kreuzkorrelation des Quellsignals mit dem gestörten Signals gekennzeichnet. Im Frequenzbereich transformiert lautet die Lösung für den Filter dann [38]:

$$H(e^{j\omega})\cdot\Phi_{xx}(e^{j\omega})=\Phi_{xs}(e^{j\omega}). \tag{62}$$

Dabei sind nun Φ_{xx} und Φ_{xs} die spektrale Leistungsdichte des gestörten Signals und die spektrale Kreuzleistungsdichte zwischen gestörtem Signal und dem Störsignal. Wird Unkorreliertheit von Störung und Nutzsignal angenommen, darf der Ausdruck Φ_{xs} mit $\Phi_{xs} = \Phi_{ss}$ und der Ausdruck Φ_{xx} als $\Phi_{xx} = \Phi_{ss} + \Phi_{nn}$ angeschrieben werden. Damit ergibt sich der Wiener-Filter im Frequenzbereich zu [38]:

$$\begin{aligned} H(e^{j\omega})\cdot\left(\Phi_{ss}(e^{j\omega})+\Phi_{nn}(e^{j\omega})\right)&=\Phi_{ss}(e^{j\omega}),\\ \Rightarrow\quad H(e^{j\omega})&=\frac{\Phi_{ss}(e^{j\omega})}{\Phi_{ss}(e^{j\omega})+\Phi_{nn}(e^{j\omega})}=\frac{\Phi_{xx}(e^{j\omega})-\Phi_{nn}(e^{j\omega})}{\Phi_{xx}(e^{j\omega})}. \end{aligned} \tag{63}$$

Über das spektrale Signal-Störverhältnis Ξ ($e^{j\omega}$) angeschrieben lautet der Ausdruck [38]:

$$H(e^{j\omega})=\frac{\Xi(e^{j\omega})}{1+\Xi(e^{j\omega})},\quad mit\quad \Xi(e^{j\omega})=\frac{\Phi_{ss}(e^{j\omega})}{\Phi_{nn}(e^{j\omega})}=max\left\{\frac{\Phi_{xx}(e^{j\omega})}{\Phi_{nn}(e^{j\omega})}-1,0\right\}. \tag{64}$$

Zum Entstören bei instationären Stör- und Nutzsignalen ist diese Methode nicht besonders gut geeignet. Üblicher Weise wird in diesem Fall eine Kurzzeit-Form des Wiener-Filters verwendet, in welchem blockweise der spektrale Signal-Störabstand bestimmt wird, und damit eine Entstörung vorgenommen wird. Der Signal-Störabstand kann über die geschätzte spektrale Rauschleistungsdichte Φ_{nn} berechnet und jener des gestörten Signals Φ_{xx} berechnet werden. Bei der Anwendung dieser Methode können sich allerdings Artefakte ergeben, die aufgrund ihrer Schmalbandigkeit *musical noise* genannt werden und äußerst störend sind [38].

4.2.3 Subtraktion der Leistungsdichtespektren

Ein verrauschtes Signal $x[n] = s[n] + n[n]$ mit dem Nutzsignal s $[n]$ und dem Störsignal n $[n]$ kann auch mit der Subtraktion der Leistungsdichtespektren entstört werden [60][38]. Dabei nimmt man eine Störung im Bereich der Leistungsdichtespektren an:

$$\Phi_{xx}(e^{j\omega})=\Phi_{ss}(e^{j\omega})+\Phi_{nn}(e^{j\omega}). \tag{65}$$

Die Entstörung ergibt sich durch den Filter H ($e^{j\omega}$):

$$\begin{aligned} \Phi_{ss}(e^{j\omega})&=H^2(e^{j\omega})\cdot\Phi_{xx}(e^{j\omega})\\ \Rightarrow\quad H(e^{j\omega})&=\sqrt{\frac{\Phi_{ss}(e^{j\omega})}{\Phi_{xx}(e^{j\omega})}}=\sqrt{\frac{\Xi(e^{j\omega})}{1+\Xi(e^{j\omega})}},\quad mit\quad \Xi(e^{j\omega})=\frac{\Phi_{ss}(e^{j\omega})}{\Phi_{nn}(e^{j\omega})}=max\left\{\frac{\Phi_{xx}(e^{j\omega})}{\Phi_{nn}(e^{j\omega})}-1,0\right\}. \end{aligned} \tag{66}$$

Die Subtraktion der Leistungsdichtespektren entspricht der Wurzel des Wiener-Filters und besitzt daher eine etwas abgeschwächte Wirkung. Die Wirkung und die dabei auftretenden Störungen sind ansonsten allerdings ähnlich wie jene des Wiener-Filters.

4.2.4 Ephraim und Malah Spectral Subtraction Rule

Ein Zusammenhang, der bei der Entstörung verrauschter Signale ($x[n] = s[n] + n[n]$) eine bessere Wirkung wie reine Wiener-Filterung oder Spektrale Subtraktion erzielt, weil Artefakte wie *musical noise* weitgehend vermieden werden, ist die Subtraktionsregel von Ephraim und Malah (kurz EMSR). Der erster Ansatz von Ephraim und Malah [30] zur Störsignalunterdrückung stammt aus dem Jahr 1984. Darin wird eine Bestimmung spektraler Gewichte $g(m,k)$ für das Frequenzband k zum Zeitpunkt m verwendet, welche die ungestörte Signalamplitude wieder herstellen soll. Durch einen *minimum mean-square short-time spectral amplitude estimator* (MMSE-STSA, oder kurz MMSE-SA) wird die Amplitude des ungestörten Signals in jedem Frequenzband geschätzt. Es gilt dabei die Annahme, dass spektrale Größen beider Signale, also des Nutz- und Störsignals, komplexwertige Gaußverteilungen besitzen. Zudem wird angenommen, dass benachbarte Frequenzbänder einer STFT (*short time fast-fourier transform*) voneinander unabhängig[6] sind. Des weiteren wird im Modell postuliert, dass auch zeitlich keine Korrelation in den Verläufen der einzelnen Spektralamplituden besteht. Die Herleitung besteht aus dem bedingten Erwartungswert für die ungestörte Spektralamplitude unter Voraussetzung der vorliegenden gestörten Spektralamplitude. Die Herleitung des Zusammenhangs in Gleichung 60 kann in der Literatur [30] gefunden werden.

Zur Berechnung des Spektralgewichts verwendete Größen sind:

- Das *a priori* Signal-Störverhältnis $\xi(m,k)$ gibt das Verhältnis der Varianzen $\sigma_s^2(k) / \sigma_n^2(k)$ von Nutz- bzw. Störsignal an und liegt den verwendeten Gaußverteilungen zugrunde. Diese Größe kann nicht direkt bestimmt werden und muss durch spezielle Verfahren geschätzt werden (z.B.: *decision directed approach*, Gl. 85).
- Ein lokales *a posteriori* Signal-Störverhältnis $\gamma(m,k)$, oder auch *instantaneous* Signal-Störverhältnis[7] $(\gamma(m,k) - 1)$, wird direkt aus der spektralen Amplitude des gestörten Signals $|X(m, k)|$ bestimmt, und besitzt daher schnell veränderlichen Charakter. Dazu muss die Schätzung

6 *Anm.: Die Signalanalyse mit der Fast-Fourier-Transformation besitzt orthogonale Eigenvektoren.*

7 *Der Schätzwert des instantaneous Signal-Störverhältnis* $\gamma(m,k) - 1$ *folgt aus dem Ausdruck*

der spektralen Rauschleistung $\sigma_n^2(k)$ bekannt sein.

Das gesuchte Gewicht $g_{\text{MMSE-SA}}(m,k)$ zur Rekonstruktion des ungestörten Amplitudenspektrums lässt sich über eine konfluente hypergeometrische Funktion M berechnen. Diese kann aus modifizierten Besselfunktionen der Ordnung 0 und 1 zusammengesetzt werden (I_0, I_1). (siehe auch Cappé [36])

$$\xi(m,k)=\frac{\sigma_s^2(m,k)}{\sigma_n^2(m,k)}. \tag{67}$$

$$\gamma(m,k)=\frac{|X(m,k)|^2}{\sigma_n^2(m,k)}. \tag{68}$$

$$g_{MMSE-SA}(m,k)=\frac{\sqrt{\pi}}{2}\sqrt{\frac{1}{\gamma(m,k)}\frac{\xi(m,k)}{1+\xi(m,k)}}\,M\left[\gamma(m,k)\frac{\xi(m,k)}{1+\xi(m,k)}\right]. \tag{69}$$

$$\text{mit}\quad M[\theta]=e^{-\frac{\theta}{2}}\left[(1+\theta)I_0\left(\frac{\theta}{2}\right)+\theta I_1\left(\frac{\theta}{2}\right)\right] \tag{70}$$

Die weiterführende Arbeit [31] von Eprhaim und Malah beschreibt einen spektralen Gewichtungsfaktor, der aus einer Schätzung der logarithmierten Signalamplitude, also mittels *minimum mean-square error log-spectral amplitude estimator* (MMSE LOG-STSA, oder kurz MMSE-LSA), gewonnen wird. Damit soll der logarithmischen Lautstärkewahrnehmung des menschlichen Gehörs Rechnung getragen weredn. Die verwendeten Größen ξ (m,*k*) und γ (*m,k*) folgen aus oben genannten Definitionen. Dieser zweite Ansatz ergibt das spektrale Gewicht $g_{\text{MMSE-LSA}}(m,k)$:

$$g_{MMSE-LSA}(m,k)=\frac{\xi(m,k)}{1+\xi(m,k)}e^{\frac{1}{2}\int_{\nu(m,k)}^{\infty}\frac{e^{-t}}{t}dt}\quad\text{, mit}\quad \nu(m,k)=\frac{\xi(m,k)}{1+\xi(m,k)}\cdot\gamma(m,k). \tag{71}$$

Das *Exponentialintegral* im Argument der *e*-Funktion lässt sich in einer Reihenentwicklung darstellen (Bronstein [61]), zu sehen in Abbildung 42:

$$E_1(\nu(m,k))=\int_{\nu(m,k)}^{\infty}\frac{e^{-t}}{t}dt=-C-\ln(\nu(m,k))-\sum_{r=1}^{\infty}\frac{[-\nu(m,k)]^r}{r\cdot r!}, \tag{72}$$

mit C=0,577215664901... (EULER'SCHE Konstante)

In Abbildung 42 ist erkennbar, dass das Exponentialintegral E_1 für kleine Werte von ν (*m,k*) einen logarithmischen Verlauf besitzt, während die Funktion für Werte große $\nu(m,k)$ >> den konstanten

$\gamma=\frac{|X|^2}{\sigma_n^2}\simeq\frac{|S|^2+|N|^2}{\sigma_n^2}\simeq\frac{|S|^2}{\sigma_n^2}+1$ *, unter Annahme von Dekorrelation zwischen Nutz- und Störsignal.*

Wert 0 annimmt.

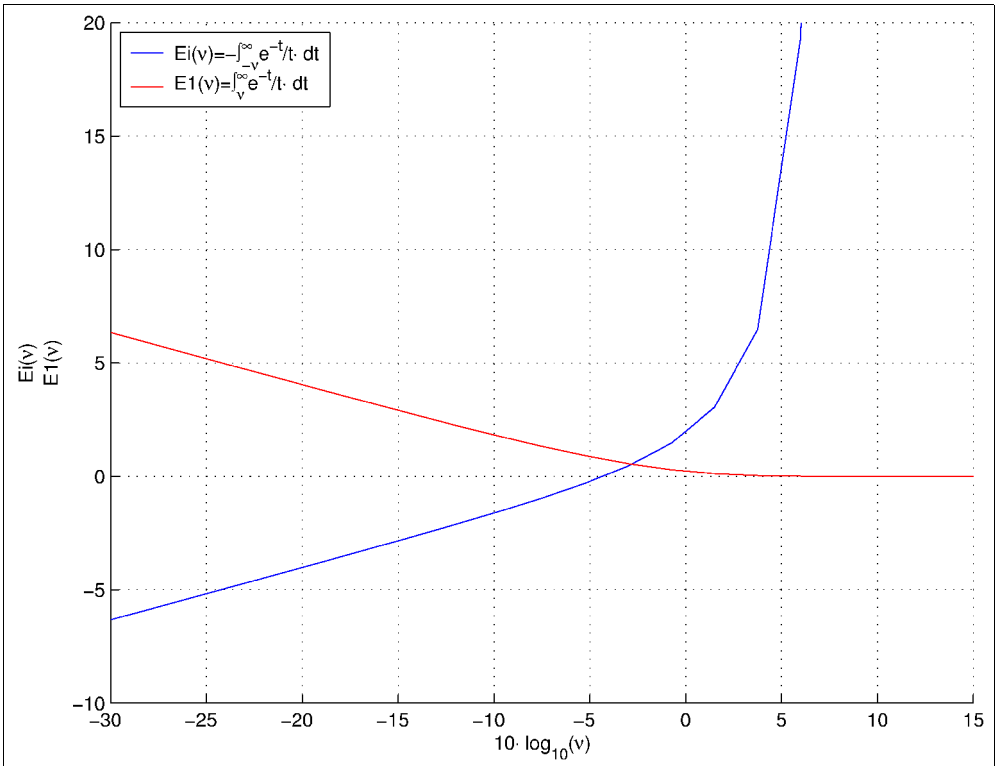

Abbildung 42: Exponentialintegral Ei(ν) und E_1(ν) in auf einer logarithmischen Skala aufgetragen.

Führt man zusätzlich den für die Wiener-Subtraktionsregel bekannten Faktor $g_{\text{wiener}}(m, k)$ ein (Abschnitt 4.2.2 und 4.2.3, siehe [62] oder [38]),

$$g_{wiener}(m,k)=\frac{\xi(m,k)}{1+\xi(m,k)}, \tag{73}$$

so kann das MMSE-LSA Spektralgewicht in einer Kurzform geschrieben werden:

$$g_{MMSE-LSA}(m,k)=g_{wiener}(m,k)\cdot e^{\frac{1}{2}\cdot E_1\left(g_{wiener}(m,k)\cdot\gamma(m,k)\right)}. \tag{74}$$

Die Form des MMSE-LSA Spektralgewichtes ist für eine praktische Implementierung geeigneter als jene des MMSE-SA Gewichtes. Den Autoren nach sind die Resultate des MMSE-LSA Gewichtungsfaktors jenen des MMSE-SA Gewichtungsfaktors trotz großer Ähnlichkeit überlegen. Um numerische Stabilität zu gewährleisten, muss für das Argument ν(*m,k*) der Wertebereich $0 < \nu\ (m,k) < \pi$ eingehalten werden. Wir schlagen vor, die Reihenentwicklung ausschließlich bei geraden Ordnungen abzubrechen. Eine Reihe bis zum Glied 6. Ordnung besitzt ausreichende

Genauigkeit.

In Gleichung 74 sieht man, dass die EMSR eine Unterdrückung mit der Methode der Wiener Spektralsubtraktion durchführt (siehe dazu auch [62] oder [38]). Dabei wirkt der multiplikativ verknüpfte Term mit der Exponentialfunktion als Korrekturterm (vgl. Cappé [36]), um das Wiener Spektralgewicht an die lokal vorgefundenen Leistungsverhältnisse anzupassen.
Es gibt zahlreiche andere Ansätze zur Rauschunterdrückung die hier nicht behandelt werden. Eine aktuelle Arbeit von Johnson, Lindgren, Povinelli und Yuan [63] beurteilt die Leistung der *Ephraim and Malah Supression Rule* (EMSR) nach wie vor als *state-of-the-art.* Die EMSR-Methode zeigt im Vergleich zu modernen Methoden des Phasenraumes sogar eine bessere Leistung. Auch in einigen jüngeren Arbeiten von Israel Cohen *et al.* [33][32][34][35] wird dem MMSE-LSA Amplitudenschätzer einige Beachtung geschenkt.

4.2.5 Approximation des Exponentialintegrals

Um die komplizierte Berechnung des Exponentialintegrals (Gl. 72) zu vereinfachen werden folgende Beobachtungen festgehalten, siehe Abbildung 48:

- Ein ganzer Term der Reihenentwicklung des Exponentialintegrals (Gl. 72) kann zur Näherung bei kleinen Argumenten $\nu << 1$ vernachlässigt werden. Die gültige Vereinfachung für das Exponentialintegral enthält den Summenterm nicht mehr $E_1(\nu) = - \ln(\nu) - C$. Die übrig gebliebenen Ausdrücke können zu $- \ln(\nu \cdot e^C)$ zusammengefasst werden. Der Definitionsbereich für diese Näherung ergibt sich bei Forderung positiver Werte ($E_1(\nu) \geq 0$) zu $\nu << e^{-C}$.
- Im Definitionsbereich $\nu >> e^{-C}$ sieht das Exponentialintegral $E_1(\nu)$ nichtmehr logarithmisch aus, sondern wie ein konstanter Ausdruck 0.

Nun könnte folgender grober Ansatz zur Approximation des Exponentialintegrals verwendet werden:

$$\begin{aligned} E_1(\nu) &\simeq max\left[-\ln(\nu)-C\,,0\right] = max\left[-\ln(\nu\cdot e^C),0\right], \\ \Rightarrow E_1(\nu) &\simeq -\ln\left[min(\nu\cdot e^C,1)\right] = -\ln\left[min(\nu\,,e^{-C})\right]-C\,. \end{aligned} \tag{75}$$

Die Begrenzung von ν in den Wertebereich $\nu > e^{-C}$ mit min (ν, e^{-C}) ergibt allerdings einen unstetigen Übergang. Eine weichere Begrenzung kann mit $\nu\, e^C / (1+ e^C)$, oder $\nu / (e^{-C}+\nu)$, beschrieben werden. Daraus resultiert eine Verfeinerung der Approximation von $E_1(\nu)$ mit:

$$E_1(\nu) \simeq -\ln\left(\frac{\nu}{e^{-C}+\nu}\right). \tag{76}$$

In der Abbildung 48 ist das Exponentialintegral im Vergleich zu den gefundenen Approximationen auf einer logarithmischen Skala aufgetragen.

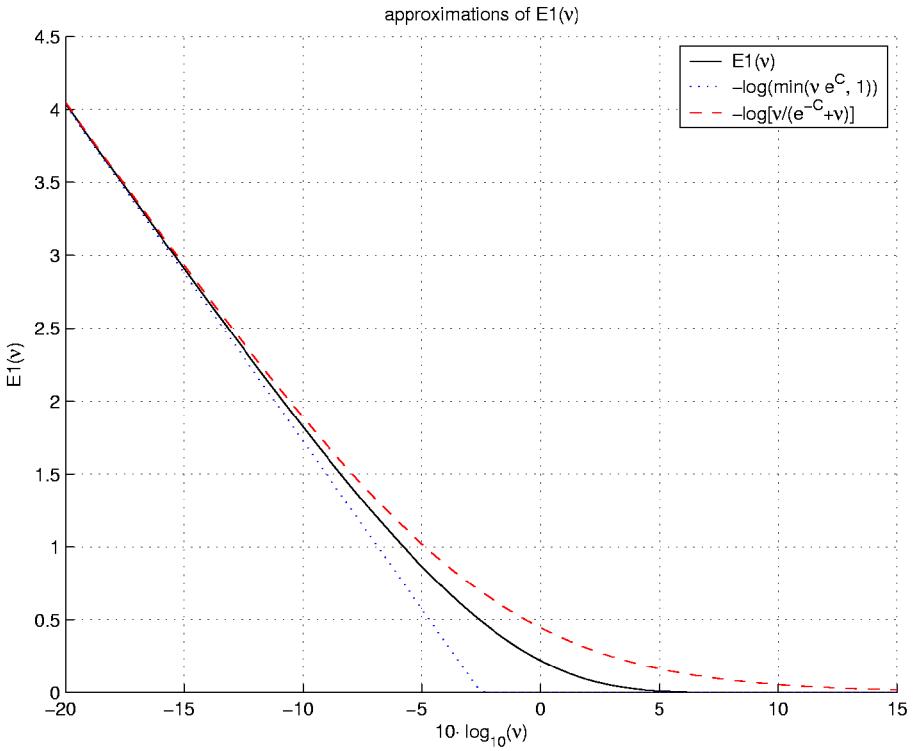

Abbildung 43: Vergleich der gefundenen Approximationen mit dem Exponentialintegral.

Eine sehr genaue Approximation konnte empirisch gefunden werden, sie bringt aber aufgrund der Exponentialfunktion keine wesentliche Vereinfachung der Berechnung mit sich:

$$E_1(\nu) \simeq -\ln\left[min\left(\nu \cdot e^{C-\frac{3{,}87\cdot\nu}{3{,}87+\nu}}, 1\right)\right]. \tag{77}$$

4.2.6 Vereinfachte EMSR Berechnung

Die Berechnung der Ephraim und Malah Subtraktionsregel kann mit der in Gleichung 76 gegebenen Approximation vereinfacht angeschrieben werden. Die zugehörige Kennfläche weicht nur sehr wenig vom Original ab, siehe Abbildung 46. Der mathematische Ausdruck ist:

$$\nu(m,k) = g_{wiener}(m,k)\cdot\gamma(m,k),$$
$$g_{MMSE-LSA}(m,k) \simeq g_{wiener}(m,k)\cdot e^{-\frac{1}{2}\cdot\ln\left[\frac{\nu(m,k)}{e^{-C}+\nu(m,k)}\right]} = g_{wiener}(m,k)\cdot\sqrt{\frac{e^{-C}+\nu(m,k)}{\nu(m,k)}}, \tag{78}$$
$$g_{MMSE-LSA}(m,k) \simeq \sqrt{g_{wiener}(m,k)\cdot\left(g_{wiener}(m,k)+\frac{e^{-C}}{\gamma(m,k)}\right)}.$$

In Wolfe und Godsill [64] wurde eine ganz ähnliche Approximation gefunden mit:

$$g_{MMSE-LSA}(m,k) \simeq \sqrt{g_{wiener}(m,k)\cdot\left(g_{wiener}(m,k)+\frac{1}{\gamma(m,k)}\right)}. \tag{79}$$

Die Approximation von Wolfe und Godsill in Gleichung 79 ist etwas ungenauer als jene in Gleichung 78. Der Unterschied beläuft sich aber nur auf den Faktor e^C, der den *a posteriori* Signal Störabstand skaliert, kann also durch entsprechende Übersubtraktion kontrolliert werden.

Ausgehend von einem der beiden Ausdrücke lässt sich die Ephraim und Malah Subtraktionsregel und der *decision directed approach* im Abschnitt 4.2.9 leichter erklären. In Martin Wittke und Jax [65] ist ebenfalls eine Näherung des Spektralgewichtes unter bestimmten Voraussetzungen zu finden.

4.2.7 Kennfläche des EMSR MMSE-LSA Spektralgewichtes

Die Subtraktionsregel von Ephraim und Malah macht das spektrale Gewicht $g_{\text{MMSE-LSA}}(m, k)$ im Wesentlichen von den zwei Parametern *a priori* SNR $\xi(m,k)$ und *a posteriori* SNR $\gamma(m,k)$ abhängig. Die Kennfläche des Spektralgewichtes $g_{\text{MMSE-LSA}}(m,k)$ ist in der Abbildung 44 zu sehen.

Mit der Vereinfachung in Gleichung 79 [64] können zwei Subtraktionsregeln gefunden werden, die im Spektralgewicht enthalten sind. Dazu wirkt der Wertebereich des Argumentes $\nu(m,k)$ der Exponentialintegral-Funktion (Gl. 72) ausschlaggebend.

Verwendet man die Approximation aus Gleichung 79, so können daraus weitere Vereinfachungen vorgenommen werden. Im Detail ergeben sich folgende numerische Ausdrücke:

$$g_{MMSE-LSA}(m,k) \simeq \sqrt{g_{wiener}(m,k)\cdot\left(\frac{\xi(m,k)}{1+\xi(m,k)}+\frac{1}{\gamma(m,k)}\right)}$$
$$A: \frac{\xi(m,k)}{1+\xi(m,k)} \ll \frac{1}{\gamma(m,k)} \Rightarrow g_{MMSE-LSA}(m,k) \simeq \sqrt{\frac{g_{wiener}(m,k)}{\gamma(m,k)}} \tag{80}$$
$$B: \frac{\xi(m,k)}{1+\xi(m,k)} \gg \frac{1}{\gamma(m,k)} \Rightarrow g_{MMSE-LSA}(m,k) \simeq g_{wiener}(m,k) \quad .$$

Die Übergangsschwelle zwischen den beiden Subtraktionsregeln *A* und *B* kann über die Identität beider Summenterme im Wurzelausdruck aus Gleichung 80 beschrieben werden. Dazu muss zwischen den beiden Parametern $\xi(m,k)$ und $\gamma(m,k)$ ein bestimmtes Verhältnis bestehen:

$$\begin{aligned} \nu(m,k) = \frac{\xi(m,k)}{1+\xi(m,k)} \cdot \gamma(m,k) = 1 \;\Rightarrow\; \gamma(m,k) &= \frac{1}{\xi(m,k)} + 1, \\ \gamma(m,k) &\ll \frac{1}{\xi(m,k)} + 1 \;\Rightarrow\; A, \\ \gamma(m,k) &\gg \frac{1}{\xi(m,k)} + 1 \;\Rightarrow\; B. \end{aligned} \tag{81}$$

Es gibt auch andere besondere Zusammenhänge:

Herrscht Einklang zwischen dem *a priori* SNR $\xi(m,k)$ und dem skalierten *instantaneous* SNR ($\gamma(m,k)$-1), ergibt sich die Subtraktion der Leistungsdichtespektren:

$$g_{MMSE-LSA}\Big|_{\gamma=\xi+1} \;\simeq\; \sqrt{\frac{\xi(m,k)}{1+\xi(m,k)} \cdot \left(\frac{\xi(m,k)+1}{1+\xi(m,k)} \right)} = \sqrt{g_{wiener}(m,k)}. \tag{82}$$

Herrscht ein reziprokes Verhältnis zwischen dem *a priori* SNR $\xi(m,k)$ und dem skalierten *instantaneous* SNR ($\gamma(m,k)$-1), ergibt sich eine gewichtete Form der Wiener Subtraktionsregel:

$$\begin{aligned} g_{MMSE-LSA}\Big|_{\gamma=1/\xi+1} \;&\simeq\; \sqrt{\frac{\xi(m,k)}{1+\xi(m,k)} \cdot \left(\frac{1}{1+1/\xi(m,k)} + \frac{\xi(m,k)}{1+\xi(m,k)} \right)}, \\ g_{MMSE-LSA}\Big|_{\gamma=1/\xi+1} \;&\simeq\; g_{wiener}(m,k) \cdot \sqrt{2}. \end{aligned} \tag{83}$$

4.2.8 Wirkungsweise der EMSR

Im Bereich *A* (Gleichung 80, Abbildung 44) verkörpert die Ephraim und Malah Subtraktionsregel die Subtraktion der Leistungsdichtespektren (Abschnitt 4.2.3) mit dem Zusatzterm Term $1/\sqrt{\gamma}$. Wird nun ein konstanter *a priori* SNR angenommen, bewirkt dieser Ausdruck, dass die spektrale Signalleistung am Ausgang der Spektralsubtraktion konstant gehalten wird. Dieses Verhalten kann *musical noise* vermindern.

$$\begin{aligned} &\underline{A \wedge \left(\xi(m,k) = \xi_{const}(k) \right) :} \\ &g_{MMSE-LSA}(m,k) = \sqrt{\frac{g_{wiener,k}}{\gamma(m,k)}}. \end{aligned} \tag{84}$$

Allerdings ist bei einer sehr großen zeitlichen Auflösung der Signalanalyse Vorsicht geboten: Wird kein Rekonstruktionsfilter zur Resynthese verwendet, so kann dieser Mechanismus als Signalverzerrung und störendes breitbandiges Rauschen wahrgenommen werden (siehe auch Abschnitt 4.2.13).

Im Bereich *B* (Gleichung 80, Abbildung 44) entspricht die Ephraim und Malah Subtraktionsregel dem Wiener-Filter (siehe auch Abschnitt 4.2.2).

Die EMSR hat also die spezielle Eigenschaft, dass bei sehr kleinen *a posteriori* Signal-Störverhältnissen eine Störgeräuschuntetrdrückung nach der Subtraktion der Leistungsdichtespektren vorgenommen wird. Dabei wird einer Verkleinerung der Signalamplitude mit dem Term $1 / \sqrt{\gamma}$ entgegengewirkt.
Zudem gilt, dass bei großem *a priori* Signal-Störverhältnissen der Wiener-Filter zur Signalentstörung zum Einsatz kommt. Dieses gegensätzliche Verhalten in Cappé [36] aufgezeigt.
Cappé [36] hat dazu eine Erklärung:

- Wenn der *instantaneous* SNR im Einklang mit dem *a priori* SNR ist (Gleichung 82), dann tritt eine Subtraktion der Leistungsdichtespektren auf. Diese kann als Arbeitspunkt der EMSR betrachtet werden
- Wennn der *instantaneous* SNR sehr viel kleiner als der *a priori* SNR ist, entspricht sein Wert nicht der statistischen Erwartung. Deshalb wird die erwartete Signalleistung am Systemausgang relativ zur verwendeten Subtraktionsregel angehoben und so wieder hergestellt.
- Wenn der *instantaneous* SNR sehr viel größer als der *a priori* SNR ist, entspricht sein Wert auch nicht der statistischen Erwartung. Deshalb kommt mit dem Wiener-Filter die stärkere Variante der Störgeräuschunterdrückung zum Einsatz.

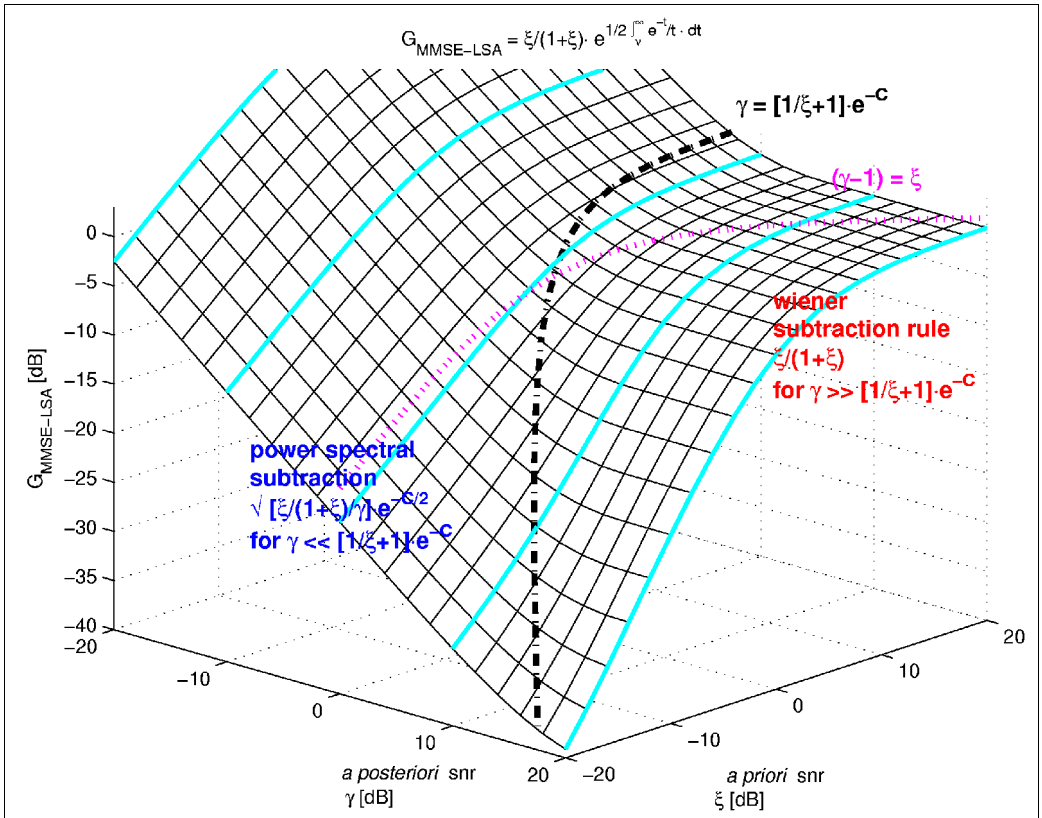

Abbildung 44: EMSR MMSE-LSA Kennfläche. Das spektrale Gewicht ergibt sich einerseits aus dem langsam veränderlichen a priori SNR und andererseits dem a posteriori SNR. Der schnell veränderliche a posteriori SNR hilft dabei musical noise lokaler Störungsspitzen zu unterdrücken, indem das Spektralgewicht zwischen der schwächeren „power spectral subtraction" und der stärkeren „Wiener"-Subtraktionsregel variiert wird (vgl. [30]).

4.2.9 Schätzung des *a priori* Signal-Störverhältnisses durch den *decision directed approach*

In der ersteren Arbeit von Ephraim und Malah [30] wird der *decision directed approach* Algorithmus zur Schätzung des *a priori* Signal-Störverhältnisses $\xi(m,k)$ vorgestellt. Hier wird $\xi(m,k)$ über den verzögerten *SNR* aus geschätzter Nutzsignalleistung zu geschätzter Störleistung und über das *instantaneous* Signal-Störverhältnis $(\gamma(m,k) - 1)$ gewonnen. β beschreibt eine Art Mittelungskonstante. Dieser Algorithmus zur Schätzung des *a priori* SNRs $\xi(m,k)$ wird auch in der Patentschrift [3] und in der Arbeit von Cornelia Falch über dieses Patent [2] verwendet. Er ist ein wichtiger Bestandteil der Subtraktionsregel von Ephraim und Malah (EMSR):

$$\xi(m,k) = \beta \frac{|Y(m-1,k)|^2}{\sigma_n^2(m-1,k)^2} + (1-\beta)\,max\left(\gamma(m,k)-1,0\right),\ mit \tag{85}$$

$$Y(m,k) = g_{MMSE-LSA}(m,k) \cdot X(m,k). \tag{86}$$

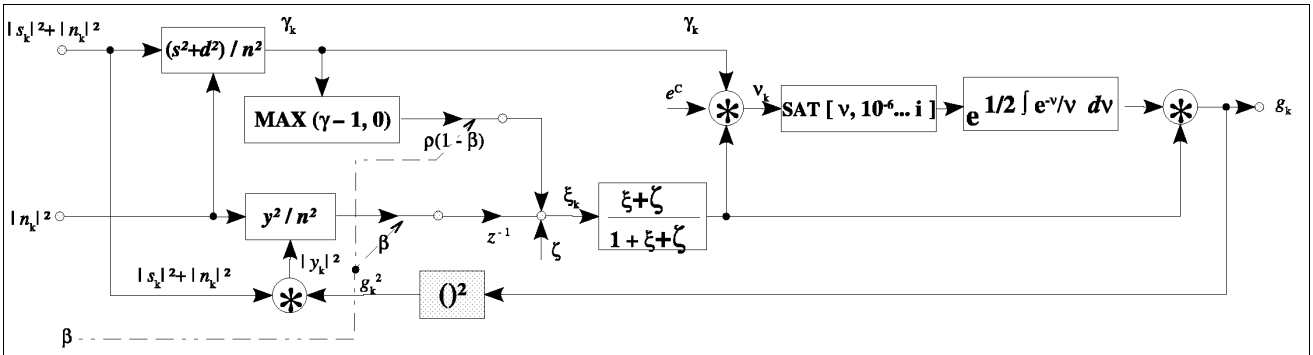

Abbildung 45: Berechnung eines spektralen Gewichts nach der MMSE-LSA Methode mit dem decision directed approach zur Bestimmung des a priori SNR im Frequenzband k.

Das *a posteriori* Signal-Störverhältnis $\gamma(m,k)$ bleibt normalerweise im Wertebereich über 1. Um dies auch in Sonderfällen zu erzwingen, muss in der Berechnung des *instantaneous* Signal-Störverhältnis eine Begrenzung $\gamma(m,k) \geq 1$ eingeführt werden, damit negative Werte für dieses Signal-Störverhältnis verhindert werden.

4.2.10 Erklärung der Funktionsweise des *decision directed approach*

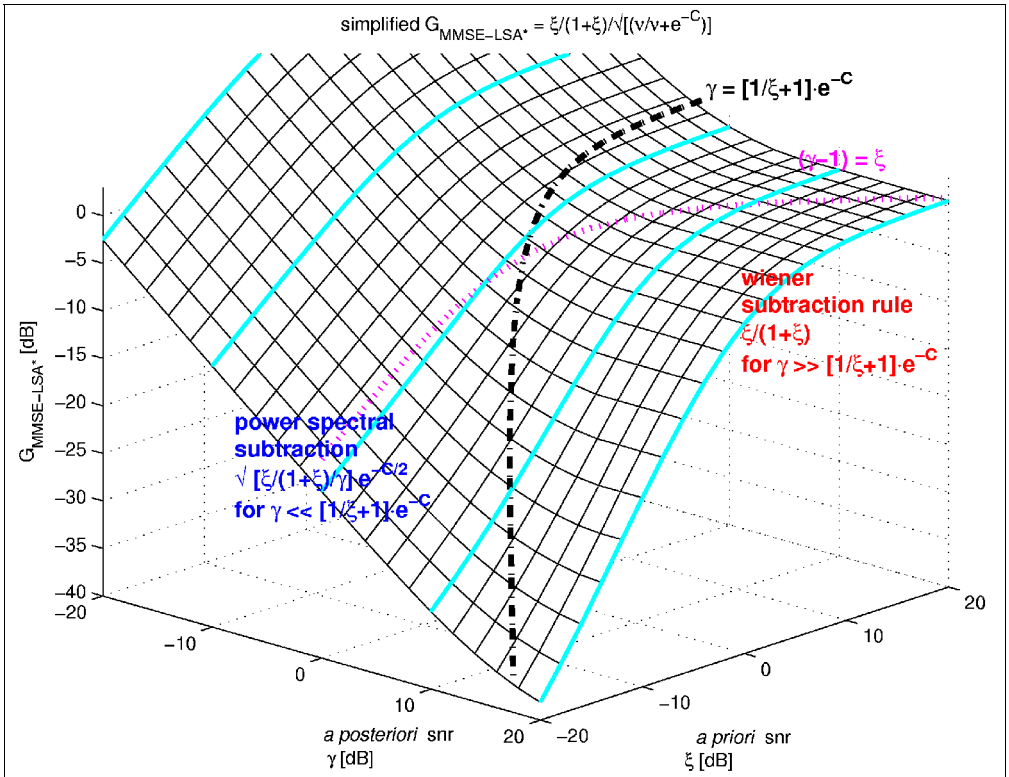

Abbildung 46: Die Kennfläche der vereinfachten EMSR (Gl. 78). Die Abweichungen des Gains von der ursprünglichen Funktion (Gl. 74) liegen in einem Bereich von 1 dB.

Der ursprüngliche *decision directed approach* von Ephraim und Malah [30] ist in Gleichung 85 beschrieben. Setzt man Gleichung 86 in Gleichung 85 ein, und fügt man die von Olivier Cappé [36] vorgeschlagene Begrenzung des *a priori* SNR ein, erhält man den Ausdruck:

$$\xi(m,k) = max\left\{\beta\cdot g^2_{MMSE-LSA}(m-1,k)\cdot\gamma(m-1,k)+(1-\beta)\cdot max\left(\gamma(m,k)-1,0\right),\xi_{min}\right\}. \qquad (87)$$

Die Wirkungsweise der EMSR zusammen mit dem *decision directed approach* wird in Olivier Cappé [36] und Israel Cohen [34] erklärt. Für kleine Signal-Störverhältnisse wird der Schätzwert für den *a priori* SNR aus einer rekursiven Mittelung des *instantaneous* SNR gebildet. Beim Übergang zu großen *instantaneous* Signal-Störverhältnissen steigt der *a priori* SNR bis in die Nähe des Wertes 1 langsam an. Danach folgt der *a priori* SNR ohne Mittelung dem um ein Abtastintervall verzögerten *instantaneous* SNR. Für große Werte von $0 < \beta < 1$ ergibt sich eine starke Verschmierung transienter Signalansätze. In Abbildung 47 wird ein Mechanismus gezeigt, der dieser Verschmierung entgegenwirken kann.

Wir versuchen eine Erklärung der Verhaltensweise des *decision directed approach* über das vereinfachte Spektralgewicht aus Gleichung 79 von Wolfe und Godsill [64] zu finden:
Nach den Gleichung 80 gibt es 2 Teilbereiche des MMSE-LSA Spektralgewichtes auf der Kennfläche. Die dabei auftretenden Subtraktionsregeln können in Gl. 85 eigesetzt werden (*Annahme:* $\xi < \xi_{min}$):

1. Wiener Subtraktionsregel:

$$g_{MMSE-LSA}\Big|_{\gamma \gg \left(\frac{1}{\xi}+1\right)} \simeq g_{wiener}(m,k)$$

Setzt man den Ausdruck $g_{\text{MMSE-LSA}}(m\text{-}1,k)$ in die Gleichung 85 ein, erhält man:

$$\xi(m,k) = \beta \cdot g^2_{wiener}(m-1,k) + (1-\beta) \cdot max\left(\gamma(m,k)-1,0\right). \tag{88}$$

Dabei ergeben sich 2 Fälle für $\xi\,(m,k)$:

a) $\xi << 1, \gamma >> 1 / \xi$:

$$\xi(m,k) = \beta \cdot \xi^2(m-1,k) \cdot \gamma(m-1,k) + (1-\beta) \cdot \left(\gamma(m,k)-1\right) \simeq (1-\beta) \cdot \left(\gamma(m,k)-1\right). \tag{89}$$

b) $\xi >> 1, \gamma >> 1$:

$$\xi(m,k) = \beta \cdot \gamma(m-1,k) + (1-\beta) \cdot \left(\gamma(m,k)-1\right) \simeq \beta \cdot \gamma(m-1,k). \tag{90}$$

2. Subtraktion der Leistungsdichtespektren:

$$g_{MMSE-LSA}(m,k)\Big|_{\gamma \ll \left(\frac{1}{\xi}+1\right)} \simeq \sqrt{\frac{g_{wiener}(m,k)}{\gamma(m,k)}}$$

Setzt man den Ausdruck $g_{\text{MMSE-LSA}}(m\text{-}1,k)$ in die Gleichung 85 ein, erhält man:

$$\xi(m,k) = \beta \cdot g_{wiener}(m-1,k) + (1-\beta) \cdot max\left(\gamma(m,k)-1,0\right). \tag{91}$$

Daraus ergeben sich 2 Fälle für $\xi\,(m,k)$:

a) $\xi << 1, \gamma << 1 / \xi$:

$$\xi(m,k) = \beta \cdot \xi(m-1,k) + (1-\beta) \cdot max\left(\gamma(m,k)-1,0\right). \tag{92}$$

b) $\xi >> 1, \gamma << 1$:

kommt praktisch nicht vor.

3. Bei Einklang zwischen *instantaneous* und *a priori* SNR:

$\gamma = (1+\xi)$:

$$\xi(m,k)=\beta\cdot\xi(m-1,k)+(1-\beta)\cdot max\left(\gamma(m,k)-1,0\right). \tag{93}$$

Aus oben gezeigten Zusammenhängen ist offensichtlich erkennbar, was der *decision directed approach* bewirkt. Die Gleichungen 92 und 93 besitzen die Form einer rekursiven Mittelung des *instantaneous* SNR.

Für die praktische Anwendung des *decision directed approach* wird in der Arbeit von Olivier Cappé [36] vorgeschlagen den *a priori* SNR nach unten hin etwa bis zum Wert ξ_{min} = -25 dB[8] zu begrenzen und einen Mittelungsparameter $\beta \approx 0{,}98$ zu verwenden (vgl. Abschnitt 4.2.12).

Obwohl in diesem Algorithmus rekursive Mittelungen auftreten, lässt sich der Mittelungsparameter β nicht über eine Zeitkonstante beschreiben. Das liegt daran, dass in dem Bereich, in welchem die Gleichung 89 Gültigkeit besitzt, eine Dämpfung (1-β) zum Einsatz kommt, die nicht für alle Abtastraten gleich groß ist. Deshalb gibt es für den Algorithmus keine Beschreibung, die allgemein für alle Abtastfrequenzen gleiches Verhalten besitzt.

Werden andere Berechnungsvorschriften als jene von Wolfe und Godsill [64] (Gl. 79) gewählt, z.B. Gl. 74 oder Gl. 78, so ist es unbedingt nötig, den *a posteriori* SNR speziell zur Bildung des Spektralgewichtes mit $\gamma\, e^{C}$ zu überschätzen, um die selbe Verhaltensweise zu erzielen. Der *decision directed approach* sollte dabei dennoch mit unbelassenem Term (γ -1) arbeiten.

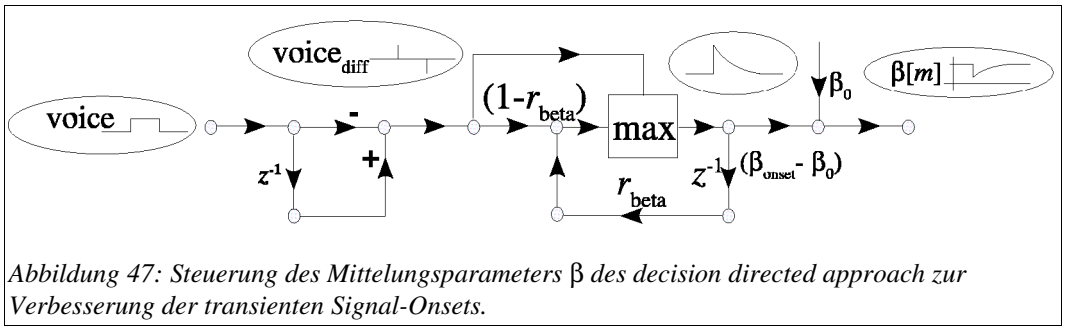

Abbildung 47: Steuerung des Mittelungsparameters β des decision directed approach zur Verbesserung der transienten Signal-Onsets.

Zur Verbesserung transienter Signalereignisse kann eine leichte Korrektur des Mittelungsparameters vorgenommen werden (vgl. Bitwave [3] und Falch [2]):

Der Mittelungsparameter $\beta[m]$ kann kurzfristig so verkleinert werden, dass während Signal-Onsets der rasch veränderliche *a posteriori* SNR $\gamma(m,k)$ stärker in die Schätzung des Signal-

8 *Anm.: Dieser Wert wird für eine FFT-Größe von etwa 256 vorgeschlagen. Werden nur 20 Frequenzbänder verwendet, muss diese Begrenzung auf etwa -25dB-10log$_{10}$(256/20) gesenkt werden um den selben Dynamikbereich zu erhalten.*

Störverhältnisses $\xi(m,k)$ eingeht. Dadurch kann sich das spektrale Gewicht $g_{\text{MMSE-LSA}}(m,k)$ rasch an das neue Signal-Störverhältnis anpassen und ein transientes Signal durchlassen. Diese kurzfristige Verkleinerung des zeitveränderlichen $\beta[m]$ kann entweder aus der globalen Sprach-Sprachpausenerkennung (VAD) oder aus dem Vergleich mit der lokalen Schwelle $a_k^2[m] > \text{Tx}_k[m]$ abgeleitet werden. Dabei muss darauf geachtet werden, dass durch diese Korrektur kein *musical noise* hervorgerufen wird. In Abbildung 47 werden beide Kriterien der Einfachheit wegen *„voice“* genannt.

Um auch hier die Eigenschaften von transienten Signalansätzen möglichst ans menschliche Gehör anzupassen, schlagen wir vor, die Mittelungsparameter zur Mittelung von $\beta[m]$ zwischen β_{onset} (Signalansatz) und β_0 (sonst) mit der Zeitkonstante der Vorverdeckung des Gehörs, also etwa 3 ms, anzusetzen.

Um einen Anhaltspunkt zu geben werden mittlere Zeitkonstanten mit etwa 700 ms für β_0

20 ms für β_{onset} vorgeschlagen.

4.2.11 Übersubtraktion

In Situationen mit instationären oder impulshaltigen Störungen, kann selbst bei der Verwendung der Ephraim und Malah Subtraktionsregel nur durch Überschätzen des Störsignalpegels (Übersubtraktion) eine Reduktion von *musical noise* erreicht werden (vgl. Berouti *et al* [51]). Das entspricht einer Unterschätzung des *a posteriori* Signal-Störverhältnisses γ. Faktoren zur Übersubtraktion liegen im Bereich $1.5 < c_{\text{oversub}} < 4$.

$$\gamma(m,k) = \frac{\gamma(m,k)}{c_{oversub}}. \tag{94}$$

Speziell im angestrebten Anwendungsfall mit Hintergrundgeräuschen im Fahrzeug ist eine Übersubtraktion hilfreich, um *musical noise* zu vermeiden.

$$SNR_{inst} = max\left[\frac{\gamma(m,k)}{c_{oversub}} - 1,0\right]. \tag{95}$$

Eine Übersubtraktion bewirkt, dass nur *a posteriori* Signal-Störverhältnisse $\gamma > c_{\text{oversub}}$ zu *instantaneous* Signal-Störverhältnissen größer als 0 führen können. Damit wird eine Schwelle geschaffen, die vom Signal eine gewisse Zeit lang überschritten werden muss, um die

Störsignalunterdrückung überwinden zu können.

4.2.12 Begrenzung des Spektralgewichts, *noisefloor vs.* perfekte Entstörung

Da laut Gustafsson, Jax und Vary [58] („preserving background noise characteristics“) eine perfekte Entstörung unnatürlich klingende Fehler produzieren kann, wird in ihrer Arbeit empfohlen von einer perfekten Entstörung abzugehen. (siehe auch Virag [56]) Zur Erhaltung der Störgeräuschcharakteristika wird ein *noisefloor*-Parameter ζ eingeführt. Dieser kleine Offset wird in der Arbeit von Gustafsson *et al.* zum Spektralgewicht hinzuaddiert. Um die Regulierung durch die EMSR (Abschnitt 4.2.8) zu erhalten, kann der *noisefloor*-Parameter ζ in die Rückkopplungsschleife des *decision directed approach* eingebunden werden. Im Prinzip entspricht die Wirkung etwa dem Vorschlag von Olivier Cappé [36], den *a priori* SNR mit ξ_{min} zu begrenzen (Gl. 87, Gl. 96).
Wird die Sprachpräsenz berücksichtigt (siehe Abschnitt 4.2.15), so kann diese Modifikation entfallen.

4.2.13 Eigene Modifikationen des *decision directed approach*

Wir erlauben uns den Ausdruck aus Gl. 87 umzuschreiben, um den Parameter ρ und den *noisefloor*-Parameter ζ einzuführen:

$$\xi(m,k)=\left\{\beta\cdot g^2_{MMSE-LSA}(m-1,k)\cdot\gamma(m-1,k)+\rho\cdot(1-\beta)\cdot max\left(\gamma(m,k)-1,0\right)\right\}+\zeta. \tag{96}$$

In den 4 Teilbereichen der MMSE-LSA Kennfläche lassen sich nun bestimmten Zusammenhänge bei der Schätzung des *a priori* SNR erkennen, vgl. Abbildung 46:

1. $\xi << 1, \gamma >> 1 / \xi$:

$$\xi(m,k)=\beta\cdot\xi^2(m-1,k)\cdot\gamma(m-1,k)+(1-\beta)\cdot\rho\cdot\left(\gamma(m,k)-1\right)\simeq(1-\beta)\cdot\rho\cdot\left(\gamma(m,k)-1\right). \tag{97}$$

2. $\xi >> 1, \gamma >> 1$:

$$\xi(m,k)=\beta\cdot\gamma(m-1,k)+(1-\beta)\cdot\rho\cdot\left(\gamma(m,k)-1\right)\simeq\beta\cdot\gamma(m-1,k). \tag{98}$$

3. $\xi << 1, \gamma << 1 / \xi$:

$$\xi(m,k)=\beta\cdot\xi(m-1,k)+(1-\beta)\cdot\rho\cdot max\left(\gamma(m,k)-1,0\right). \tag{99}$$

4. $\xi >> 1, \gamma << 1$:

$$\xi(m,k)=\beta. \tag{100}$$

Auf diese Weise kann die in Olivier Cappé [36] beschriebene Fuktionsweise des *decision directed approach* noch erweitert werden. Der zusätzliche Koeffizient ρ dient zur Steuerung der Ausdehnung der Hystereseschleife (vgl. Abbildung 46). Wäre der Parameter β sehr klein und ρ=1, könnte sich keine Hystereseschleife ausbilden. Weil der neu eingeführte Faktor unabhängig von der Abtastfrequenz ist und den direkten Einfuss auf die Hysterese ermöglicht, ist die Verwendung des modifizierten *decision directed approach* gerechtfertigt.

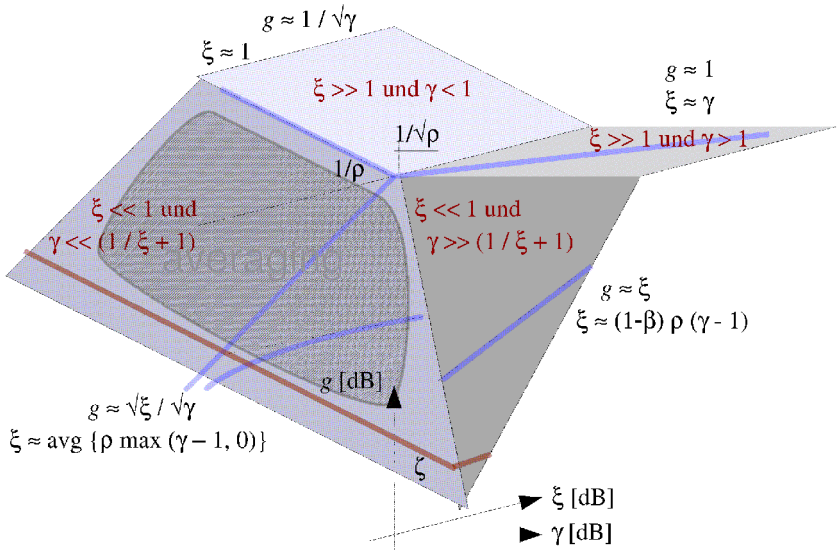

Abbildung 48: Schematische MMSE-LSA Kennfläche mit den Kennlinien des modifizierten decision directed approach zur Bestimmung des a priori SNR. Im Teilbereich in welchem die Subtraktion der Leistungsdichtespektren zu Stande kommt, erfolgt eine Mittelung des instantaneous SNR. Alle anderen Teilbereiche besitzen statische Kennlinien zur Berechnung des a priori SNR.

Wie in Abbildung 49 gezeigt wird, ergibt sich bei der Verwendung des veränderten *decision directed approach* eine breite Hystereseschleife auf der MMSE-LSA Kennfläche. Die Breite und Höhe der Hystereseschleife kann mit dem Parameter ρ gesteuert werden (Abbildung 46). Der Mittelungsparameter β hat eine etwas kompliziertere Aufgabe. Im Wesentlichen bewirkt der Mittelungsparameter eine Reduktion der Geschwindigkeit, mit der sich das *a priori* Signal-Störverhältnis ξ im Bereich der Subtraktion der Leistungsdichtespektren verändern kann, damit wird die Varianz der Schätzung bei großen Mittelungsparametern verringert und *musical noise* verringert. Im Bereich (Gl. 97) bewirkt allerdings der Faktor (1- β) bei großer Mittelungskonstante eine starke Abschwächung des SNR-Schätzwertes. An diesem Bereich ist ein eher kleiner Mittelungsparameter

wünschenswert, um eine nachvollziehbare Wirkungsweise zu erreichen.

Durch die Mittelung wird auch die Geschwindigkeit eingeschränkt, mit welcher der Wert des Spektralgewichts beim Eintritt in schlechtere Signal-Störverhältnisse verringert wird. Zu starke Mittelung verhindert beim Durchlaufen der Hystereseschleife, dass eine ausreichende Unterdrückung zu Stande kommt. Das *a priori* Signal-Störverhältnis muss zu diesem Zweck in kurzer Zeit weit genug sinken können (vgl. Abbildung 49).

Eine zu große Mittelung kann also bewirken, dass die Störsignalunterdrückung versagt.

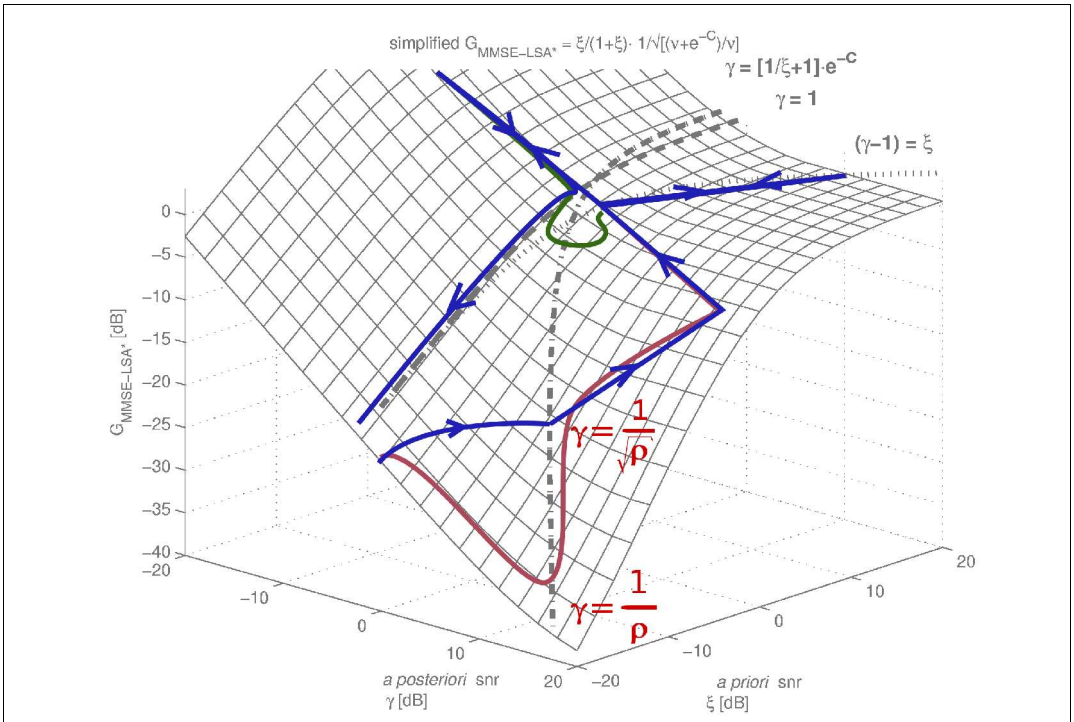

Abbildung 49: Bei der Bestimmung des a priori SNR wird eine Hystereseschleife durchlaufen, ein Beispiel ist durch die dicke blaue Linie mit den Pfeilen angedeutet. Der Parameter ρ legt die Breite und Höhe der Schleife fest. Der Mittelungsparameter β erfüllt mehrere Aufgaben. Zum Einen ist die Mittelung dafür verantwortlich, wie schnell eine Verkleinerung des Spektralgewichtes g erfolgen kann. Ist die Mittelungskonstante zu groß, versagt die Störsignalunterdrückung (grüne Linie mit zu kleiner Hysterese). Ein großer Mittelungsparameter ermöglicht aber auch eine bessere Unterdrückung von musical noise.

Des Weiteren kann sich durch den Faktor $1 / \sqrt{\gamma}$ im Spektralgewicht $g = \sqrt{(g_{wiener}/\gamma)}$ für $\gamma < (1/\xi+1)$ ein unerwünschter Effekt ergeben. Die auditive Signalanalyse besitzt eine sehr gute Zeitauflösung. Kommt nun die Korrekturwirkung des Faktors $1 / \sqrt{\gamma}$ zu Stande, wird die Signalamplitude am Ausgang der Spektralsubtraktion konstant gehalten. Dies kann starke harmonische Verzerrungen

und Störungen bewirken, welche als störend empfunden werden. Wird kein Rekonstruktionsfilter zur Resynthese eingesetzt, bleibt das Ausgangssignal gestört. Als Abhilfe dazu kann die Mittelungskonstante β klein genug gehalten werden, oder das Verfahren von Cohen angewendet werden, das die Wahrscheinlichkeit der Sprachpräsenz berücksichtigt (siehe Abschnitt 4.2.15).

In den Simulationen für den modifizierten *decision directed approach* zeigte sich, dass Zeitkonstanten von weniger als 1 ms ausreichen.

L (downsampling)	β_0	ρ
1	0,13 ms	12 dB
2	0,27 ms	11,5 dB
4	0,57 ms	11 dB
8	1,14 ms	11 dB
16	2,5 ms	10 dB
32	5,3 ms	9,5 dB
64	5,3 ms	9 dB
128	5,3 ms	9 dB

Tabelle 1: In der Simulation ermittelte Parameter für den modifizierten decision directed approach

Um die im Bereich $\xi(m,k) = (1-\beta)\,\rho\,\min(\gamma(m,k)-1,0)$ auftretende starke Dämpfung und daraus folgende Verschmierung von Transienten zu vermeiden, kann der Ansatz in Abbildung 47 verwendet werden.

Für die modifizierte Form des *decision directed approach* sind statische Mittelungsparameter im Bereich einer Milisekunde nötig, siehe Tabelle 1. Die Mittelung des Übergangs von β_{onset} zu β_0 kann mit etwa einmal der Zeitkonstante der Nachverdeckung gerechnet werden.

4.2.14 Modifikationen des *decision directed approach* von Cohen

In einer aktuellen Arbeit von Israel Cohen [34][35] werden zwei Modifikationen des *decision directed approach* zur Schätzung des *a priori* Signal-Störverhältnisses beschrieben. Zum einen wird ein akausaler Ansatz vorgeschlagen, der einige *frames* Zeitverzögerung benötigt. Dieser Ansatz soll speziell kurze Störungen unterdrücken können. Diese Veriante hier aufgrund der damit einhergehenden Verzögerung nicht verwendet. Nennenswert ist aber auch der zweite kausale Zusammenhang, den Cohen zur Schätzung gefunden hat. Dabei setzt sich die Schätzung des *a priori* SNR aus zwei Schritten zusammen:

1. „propagation“ Schritt:

$$\xi_p(m-1,k)=max\left\{(1-\beta)\cdot\xi(m-1,k)+\beta\cdot g^2_{MMSE-LSA}(m-1,k)\cdot\gamma(m-1,k),\xi_{min}\right\}. \tag{101}$$

2. „update“ Schritt:

$$\xi(m,k)=\frac{\xi_p(m-1,k)}{1+\xi_p(m-1,k)}\cdot\left(1+\frac{\xi_p(m-1,k)}{1+\xi_p(m-1,k)}\cdot\gamma(m,k)\right). \tag{102}$$

Die so erhaltene Schätzung besitzt kleinere Varianzen bei schlechten Signal-Störverhältnissen als jene des ursprünglichen *decision directed approach* [35].

4.2.15 Berücksichtigung der Sprachpräsenz

In Malah, Cox und Accardi [66] wird die Wahrscheinlichkeit für Sprachpräsenz als Modifikation des MMSE-LSA Gewichts (Gl. 74) zum MM-LSA (*modified log-spectral amplitude estimator*) beschrieben. Auch in Martin Wittke und Jax [65] wird diese Wahrscheinlichkeit zur Modifikation verwendet, um eine verbesserte Gewichtung zur Störgeräuschreduktion zu erhalten. Cohen und Berdugo verfeinern die Gewichtungsregel zum (*optimally modified*) OMM-LSA [33]. Begründet wird dieser zusätzliche Aufwand damit, dass die Herleitung der EMSR-Regel (Abschnitt 4.2.4, [30] [31]) auf dem Vorhandensein von Sprache beruht:

$$\begin{gathered} g_{OM-LSA}(m,k)=g^{p}_{MMSE-LSA}\cdot g^{(1-p)}_{min}, \\ g_{min}=const. \end{gathered} \tag{103}$$

In Cohen und Berdugo [33] und Martin *et al.* [65] wird berichtet, dass so eine Modifikation des Ephraim und Malah Spektralgewichtes in Verbindung mit dem *decision directed approach* zu einer Form von strukturiertem Rauschen führen kann. Deshalb wird zuerst ein unverändertes Spektralgewicht nach Ephraim und Malah berechnet. Erst dann wird dieses Gewicht modifiziert.

In der Arbeit von Cohen dient die *a priori* Wahrscheinlichkeit für die Sprachpräsenz *p* als Paremeter, mit welchem bei einer geometrischen Mittelung einer konstanten Dämpfung $g_{\min}$ und dem MMSE-LSA Gewicht $g_{\text{MMSE-LSA}}$ jeweils eines beider Gewichte begünstigt wird. Zur Bestimmung dieser Wahrscheinlichkeit *p* werden 3 unterschiedlich gemittelte Versionen des *a priori* SNR verwendet. Die drei Mittelungsarten umfassen eine Kurzzeit-, eine Langzeitmittelung und eine über alle Frequenzbänder. Alle drei Parameter werden umgeformt und multiplikativ verknüpft. Aus der Differenz von 1 und dem Wert dieser Multiplikation ergibt sich die Wahrscheinlichkeit *q* für Sprachabsenz. Die *a priori* Wahrscheinlichkeit der Sprachpräsenz *p* kann

daraus so berechnet werden:

$$p(m,k)=\frac{1}{1+\frac{q(m,k)}{1-q(m,k)}\cdot\left(1+\xi(m,k)\right)\cdot e^{-\nu(m,k)}}. \tag{104}$$

Dieses Verfahren bringt für die Integration in eine auditive Signalanalyse folgende Vorteile mit sich:

- In vorangegangenen Abschnitten genannte Probleme des Korrekturterms $1 / \sqrt{\gamma}$ in Verbindung mit hoher zeitlicher Analyseauflösung wird hier bei schlechter Sprachwahrscheinlichkeit abgeschächt. Somit behält das Störgeräusch seine Form und wird konstant gedämpft, ohne dass zusätzliche Störungen erzeugt werden können (siehe Ende des Abschnitts 4.2.13 und Abschnitt 4.2.8).
- *Musical Noise* wird verringert, indem ein konstanter *noisefloor* eingeführt wird. Damit besitzt die Spektrale Subtraktion die Eigenschaft keine perfekte Entstörung mehr durchzuführen. Unnatürlich klingende Artefakte werden somit weiter vermieden, indem das Störgeräusch noch erkennbar zu hören ist.

Aufgrund des erheblichen Berechnungsaufwandes wird diese interessante Gewichtungsregel allerdings in unserer Implementation nicht verwendet.

Eine weitere interessante Arbeit von Cohen verwendet in einem System mit Mikrofonzeile einen Generalized Sidelobe-Canceller zur Bestimmung der Wahrscheinlichkeit für die Absenz von Sprache [67].

4.3 Ressourcenaufwand einer vollständigen auditiven Spektralanalyse

Ein Außen-Mittelor-Übertragungsfilter und das inverse Filter benötigen

	Mult	***Add***	***Coeffs***	***Storage***
HP	4	4	2	2
HP+RES	9	8	7	4
HP+RES+TP	13	12	9	6

Eine Gammatone-Filterbank mit N Kanälen und der Ordnung M benötigt

	Mult	***Add***	***Coeffs***	***Storage***
APGF	3 N M	2 N M	3 N	2 N M

	Mult	***Add***	***Coeffs***	***Storage***
OZGF	3 N M	2 N M + 1	3 N	2 N M + 1
TZGF	3 N (M+1) - N	2 N (M+1) + 1	5 N	2 N M + 1
GF	4 N M	3 N M	3 N + N M	2 N M

Bei N_{IQ} Kanälen benötigt eine 90° Phasendrehung (für die Pegelbestimmung bei tiefen Frequenzen)

	Mult	***Add***	***Coeffs***	***Storage***
FIR 1^{st} Order	2 N_{IQ}	N_{IQ}	2 N_{IQ}	N_{IQ}
Allpass 1^{st} Order	2 N_{IQ}	2 N_{IQ}	2 N_{IQ}	N_{IQ}
Delay				jeweils bestimmen!

Bei N_{IQ} Kanälen benötigt eine Pegelbestimmung mittels Phasenschieber bei einem Downsampling von L

	Mult	***Add***	***Coeffs***	***Storage***
$x_I^2[m]+x_Q^2[m]$	4 N_{IQ} / L	N_{IQ} / L		

Bei $N_{ABS} = (N - N_{IQ})$ Kanälen benötigt eine Pegelbestimmung mit Absolutbetragsbildung

	Mult	***Add***	***Coeffs***	***Storage***
LP{‖ $x[n]$ ‖}	2 N_{ABS} +2 N_{ABS} / L	N_{ABS}	2	N_{ABS}
21-Sample-Delay for reference sigs				21 N_{ABS}

Bei N Kanälen benötigt die Modellierung der Nachverdeckung bei einem Downsampling L

	Mult	***Add***	***Coeffs***	***Storage***
postmask	2 N / L	N / L	2 N	N

+ N / M Vergleichsoperationen

Bei N Kanälen benötigt eine inverse Gammatone-Filterbank

	Mult	***Add***	***Coeffs***	***Storage***
summation		N - 1		
summation+sign	N	N - 1	N	

Einige Varianten sind unten zu sehen, für $N_{IQ} = N_{ABS} = N / 2$

	Mult	***Add***	***Coeffs***	***Storage***
APGF, FIR-90°, summation	N (3 M + 2 + 5 / L) + 18	N (2 M +2+1,5 / L) + 15	6 N + 16	N (2 M + 23) + 8
OZGF, FIR-90°, summation	N (3 M + 2 + 5 / L) + 18	N (2 M +2+1,5 / L) +16	6 N + 16	N (2 M + 23) + 9
TZGF, FIR-90°, summation	N (3 M + 4 + 5 / L) + 18	N (2 M +4+1,5 / L) +16	8 N + 16	N (2 M + 23) + 9
GF, FIR-90°, summation	N (4 M + 2 + 5 / L) + 18	N (3 M +2+1,5 / L) + 15	N (M + 6) + 16	N (2 M + 23) + 8

Für N = 20 ergibt liegt der Rechenaufwand (Additionen+Multiplikationen) bei ca. 7,9 MIPs, wenn eine Samplingrate von 16 kHz bei Analyse mit einer TZGFB 3. Ordnung vorgegeben ist.

Eine FFT-Analyse (N=256) mit Bark-Transformation (N=20, ohne Spreading), einem Analyse- und Resynthesefenster, sowie einem 50% Overlap verwendet etwa 5,3 MIPs, wenn in reellen Größen gerechnet wird.

Der größere Rechenaufwand bringt aber auch einige zusätzliche Eigenschaften mit sich: geringe Latenzzeit, Außen-Mittelohr-Übertragugnsfunktionen

In Abbildung 50 ist die Struktur der gesamten auditiven Signalanalyse anhand eines Beispiels zu sehen.

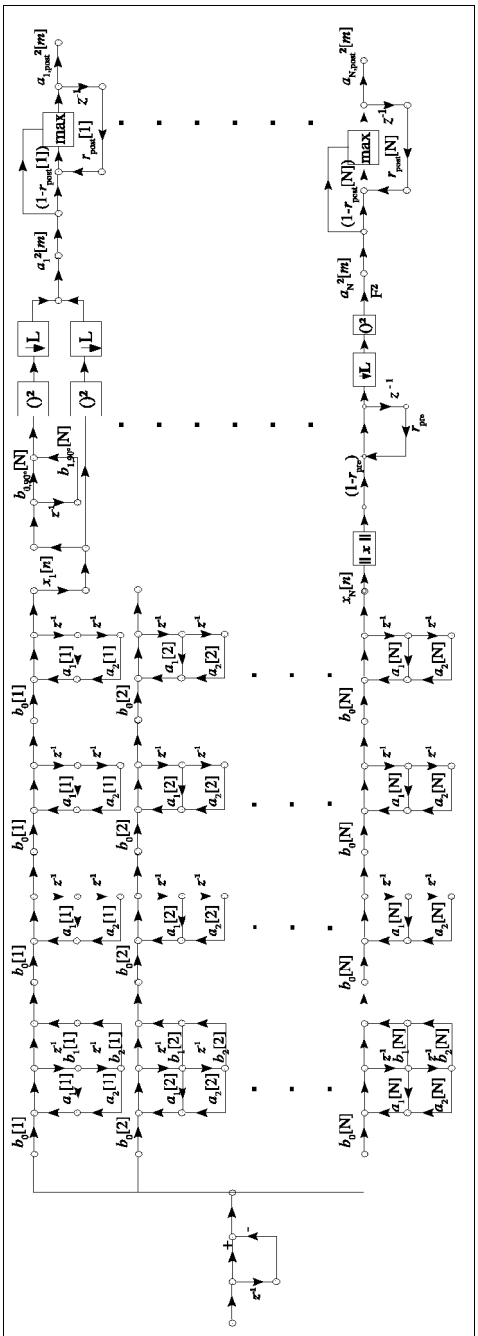

Abbildung 50: Beispiel für die Struktur einer auditiven Signalanalyse.

4.4 Ressourcenaufwand der Störsignalunterdrückung

Eine Störsignalschätzung benötigt folgende Anzahl an Operationen pro Sample, wenn L die Unterabtastung und N die Anzahl der Frequenzbänder ist:

	Mult	***Add***	***Coeffs***	***Storage***	***Compare***
with VAD (loudness)	(4+2N)/L	(2+2N-1)/L	6+N	4+N	(5+2N)/L
without VAD	6N/L	3N/L	4N+3	5N	7N/L

Die vereinfachte Berechnung des Ephraim und Malah Spektralgewichtes samt *„decision directed approach“* zur Schätzung des *a priori* SNR benötigt folgende Anzahl an Operationen pro Sample wenn L die Unterabtastung ist und N die Anzahl der Frequenzbänder:

	Mult	***Add***	***Coeffs***	***Storage***	***Squareroot***	***Division***
EMSR+DD	6N/L	3N/L	2	N	N/L	4N/L

Zum Hinauftasten des Signals auf die ursprüngliche Abtastrate benötigt man mit der linearen Interpolation pro Sample folgende Anzahl an Operationen, hier werden wieder N Samples betrachtet:

	Mult	***Add***	***Coeffs***	***Storage***
Upsampling (lin)	N/L	N-N/L	1	0

Die spektrale Gewichtung des Signals benötigt folgende Anzahl an Operationen:

	Mult	***Add***	***Coeffs***	***Storage***
Vocoder	N	0	0	0

In Summe ergibt sich als Rechenaufwand wenn keine explizite VAD zur Störsignalschätzung verwendet wird:

	Mult	***Add***	***Coeffs***	***Storage***	***Squareroot***	***Division***	***Compare***
Suppression	13N/L+N	5N/L+N	4N+6	6N	N/L	4N/L	8N/L

Das sind in Summe 31N/L+2N Operationen. Für eine Implementierung mit 20 Frequenzbändern und 32-facher Unterabtastung ergibt sich damit eine Gesamtanzahl 60 von Operationen pro Sample, also etwa 1 MIPs bei einer Samplingrate von 16 kHz.

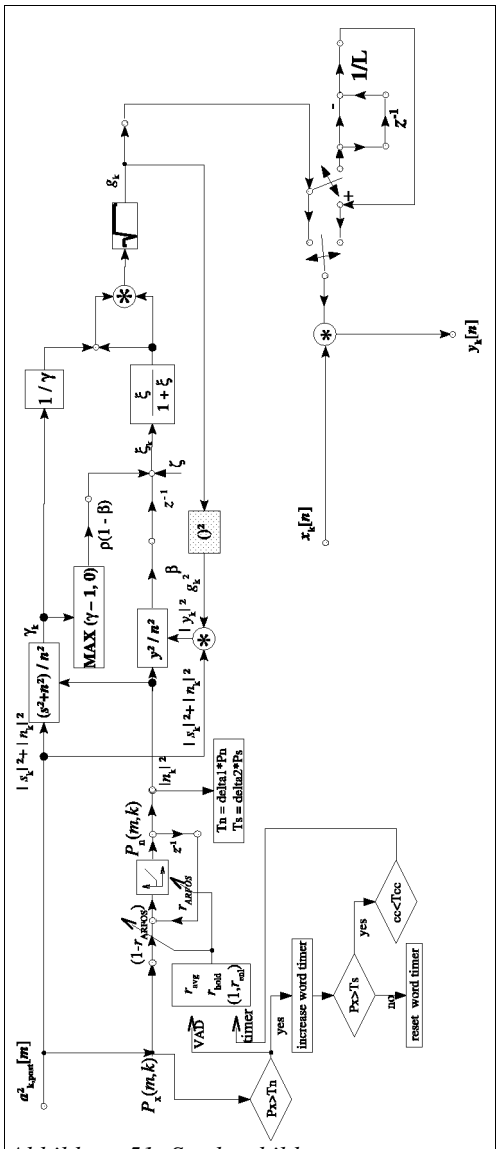

Abbildung 51: Strukturbild zur Störsignalunterdrückung.

4.5 Ressourcenaufwand einer gesamten auditiven Störgeräuschunterdrückung

Die Anzahl von Operationen pro Samples liegt für die gesamte auditive Störgeräuschunterdrückung dann bei 494+60=556. Das sind für eine 16 kHz Samplinrate 8,9 MIP/s.

(*Anm: Eine Störgeräuschunterdrückung mit FFT-Frequenzanalyse benötigt etwa 6 MIPs*)

5 Ausblick

Es werden noch weiterführende Arbeiten im Zusammenhang mit der auditiven Signalverarbeitung und dem System, welches im Einleitungskapitel erwähnt wird, durchgeführt.
Die gefundenen Systemparameter für adaptives Beamforming, und alle anderen Einstellungen für das am Institut aufgebaute System aus des Bitwave-Patent [3] und der Arbeit [2] müssen noch dokumentiert werden. Es sollen Audiobeispiele aufgenommen werden. Zur weiteren Verfeinerung sind noch einige Verbesserungen an den Algorithmen und Zusatzmaßnahmen nötig. In den folgenden Abschnitten sind die begonnenen Modifikationen und Überlegungen zu lesen.

5.1 Howling-Suppression

Eine aktive akustische Verstärkung eines Mirkofonsignals über Lautsprecher kann zu Rückkopplungen führen, wenn der Übertragungspfad Lautsprecher-Mikrofon zu wenig gedämpft wird. Bereits vor Rückkopplungen machen sich schmalbandige Signalstörungen, die als Pfeifen oder *howling* bekannt sind. Bereits bei kleinen Veränderungen des Rückkopplungspfades kann sich dann instabiles Verhalten ergeben. Dabei wird dieses Pfeifen sehr laut, und kann sogar zur Zerstörung von Komponenten des Signalwegs führen.
Um die Störungen und die Rückkopplung zu vermeiden, kann ein Algorithmus zur *howling detection* und *howling suppression* verwendet werden.
Da vor und bei Rückkopplungen meist schmalbandige, fast Sinusförmige Störungen entstehen, kann *howling* leicht durch ein Maß der spektralen Flachheit (SFM, *spectral flatness measure* [24]) erkannt werden. Ist dieses Maß sehr groß, ist das analysierte Signal rauschähnlich und besitzt ein flaches Spektrum, wenn das Maß hingegen sehr klein bzw. 0 ist, dann besitzt das analysierten Signal vereinzelte Frequenzbänder mit Signalkomponenten, während andere Frequenzbänder keinen Inhalt haben.

$$SFM[m] = 10 \cdot \log_{10}\left(\frac{G[m]}{A[m]}\right) = 10 \cdot \log_{10}\left(\frac{\sqrt[N]{\prod_{k=1}^{N} a_k^2[m]}}{\frac{1}{N}\sum_{k=1}^{N} a_k^2[m]}\right). \tag{105}$$

Wird nun über einen empirisch festgelegten Schwellwert $SFM_{howling}$ *howling* detektiert, so kann die

Unterdrückung dieses Rückkopplungsgeräuschs in die Störsignalunterdrückung integriert werden. Dazu müssen von *howling* betroffene Frequenzbänder in der Spektralschätzung des Störsignals auf „Mitteln“ geschaltet werden. Dies ist deshalb nötig, weil sich *howling* zum Teil so schnell aufbaut, dass es in der Statistik zur Schätzung stationärer Störungen nicht enthalten ist. Werden nun die betroffenen Frequenzbänder als Störung erkannt, können die Störgeräusche von der spektralen Subtraktion eliminiert und Rückkopplungen verhindert werden.

Frequenzbänder, die *howling* beinhalten können so erkannt werden:

$$a_k^2[m] > c_{howling} \cdot A[m] \quad band\ k\ contains\ howling. \tag{106}$$

Dabei wird die Konstante $c_{howling}$ so festgelegt, dass nur jene Bänder, die ausreichend weit von der mittleren Leistung abweichen als *howling* indiziert werden.

Eine Struktur zur Howling Detection könnte so aussehen wie in Abbildung 52.

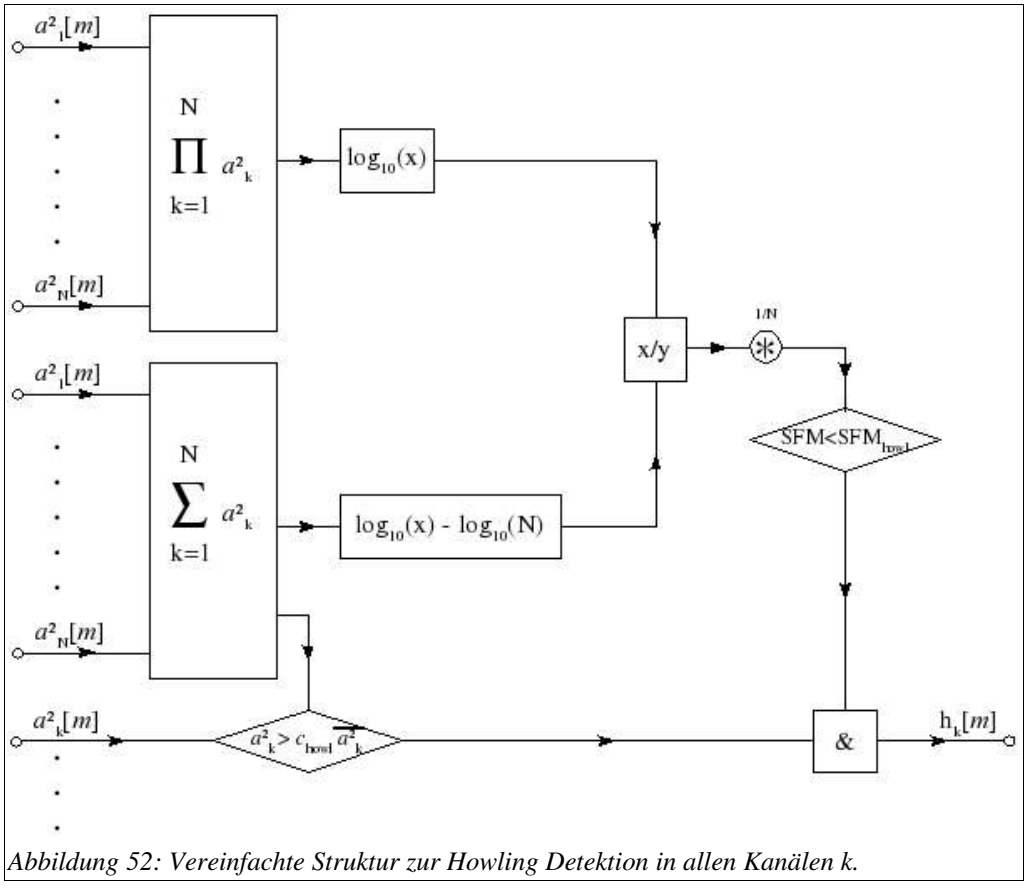

Abbildung 52: Vereinfachte Struktur zur Howling Detektion in allen Kanälen k.

Dabei soll das Signal $h_k[m]$ die *howling*-Entscheidung darstellen. Diese muss an die spektrale Störsignalschätzung so angeschlossen werden, dass ein logischer Pegel 1 die Mittelung im

Frequenzband *k* aktiviert. Auf diese Weise kann schmalbandiges *howling* effizient unterdrückt werden. Für N Kanäle besitzt diese effiziente Methode zur *howling*-Erkennung N Multiplikationen, N Additionen, 2 Logarithmen, eine Division und 2N+1 logische Operationen.

5.1.1 Mit einem LMS-Notchfilter

Zur *howling detection* und *suppression* wird oft ein LMS-Notchfilter verwendet, der schmalbandige Störungen unterdrücken kann (Haykin [62]). Dieser adaptive Filter erhält am Eingang eine Referenzschwingung, die ungefähr dieselbe Frequenz wie die zu unterdrückende Störungen besitzt. Diese Schwingung muss allerdings bereits im Voraus bekannt sein. Ist der Übertragungsweg Lautsprecher-Mikrofon bekannt und besitzt dieser einen Frequenzgang mit einigen Überhöhungen, dann sind die zu erwartenden Störfrequenzen des *howling* direkt abzulesen.
Ist die Frequenz der schmalbandigen Störung allerdings nicht bekannt, so muss eine Schätzung der Störsignalfreqenz mittels FFT oder PLL erfolgen.

5.1.2 Erkennung der Störfrequenz mit einer PLL-Schaltung

Phase-locked loop-Schaltungen werden normalerweise dazu verwendet, um Schwingungen mit derselben Frequenz phasengenau aufeinander einzustellen. Je nach Dimensionierung besitzt eine PLL-Schaltung jedoch auch die Fähigkeit in breitbandigen Signalen Signalfrequenzen einzufangen. Leider sind PLL-Schaltungen bei einem solchen Verwendungszweck empfindlich gegenüber Störungen und Rauschen, weshalb hier keine sinnvolle Konfiguration einer PLL-Schaltung zur ausreichend genauen Detektion von Störungfrequenzen gefunden werden konnte.

5.2 Echo-Cancellation/Suppression

In Telekommunikationssystemen ist eine gewisse Dämpfung vorgeschrieben, mit welcher das das empfangene Signal des fernen Sprechers im gesendeten Signal des nahen Sprechers enthalten sein darf. Freisprechanlagen können die geforderte Dämpfung nur unter Einsatz spezieller Algorithmen erreichen, welche das im Raum akustisch veränderte Signal des fernen Sprechers aus dem Signal des Mikrofones subtrahieren können. Gelingt das nicht, hört der ferne Sprecher sein Sprachsignal verzögert als Echo am eigenen Hörer. Noch schlimmer: Zwischen 2 Benutzern von Freisprechanlagen kann bei Versagen der Echo-Kompensation eine Rückkopplung entstehen [38].
Eine Kompensation akustischer Echos (*acoustic echo cancellation,* AEC) besteht im Prinzip aus

einem Filter, der den Übertragungsweg des Signals im Raum vom Lautsprecher zum Mikrofon nachbilden kann. Dabei treten folgende Schwierigkeiten auf [68]:

- Akustische Übertragungswege besitzen aufgrund der langsamen Schallausbreitung und der großen Wellenlängen des Luftschalls sehr lange Impulsantworten. Filter, die solche Übertragungswege modellieren sollen müssen etwa 256-1024 Samples lang sein [38]. Es ist also mit erheblichem Rechenaufwand zu rechnen. Für die gewünschte Echoreduktion um R_{echo} kann bei einer Nachhallzeit von T_{60} die benötigte Filterlänge abgeschätzt werden [68]: $N_{\text{AEC}} = f_{\text{s}}\, T_{60}\, R_{\text{echo}} / 60$.
- Der akustische Übertragungspfad Lautsprecher-Mikrofon bleibt nicht konstant. Bewegen sich Personen im Raum (Sprecher), oder verändern Gegenstände ihre Position im Raum, verändert sich die Impulsantwort des Übertragunspfades. Auch die Temperatur, welche die Schallgeschwindigkeit beeinflusst, kann diesen Übertragungspfad leicht verändern.

  Zur Identifikation des Echopfades muss ein adaptiver Algorithmus eingesetzt werden, der Veränderungen entsprechend schnell folgen kann und große Genauigkeit besitzt.
  - Aufgrund benötigten Filterlänge ist eine schnelle Adaption des modellierten Echopfades schwer erreichbar. Eine schnelle Adaption steht im Gegensatz zur benötigten Genauigkeit [38].
  - Da im Raum auch Störquellen vorhanden sind, sowie Umgebungsgeräusche oder der nahe Sprecher selbst, ist die Adaptionsgenauigkeit eingeschränkt. Im Idealfall werden die optimalen Filterkoeffizienten am effizientesten gefunden, wenn das Mikrofon bloß den Lautsprecher empfängt und dieser weißes Rauschen einspielt.

    Eine Aktivität des nahen Sprechers, Störgeräuschquellen und gefärbte Lautsprechersignale beeinträchtigen das Adaptionsverhalten. Meist wird eine Pegelwaage eingesetzt, welche die Pegel des Mirkofon und Lautsprechersignals vergleicht. Werden im Raum zu hohe Pegel detektiert, kann die Adaptionsgeschwindigkeit verlangsamt werden [38].
  - Werden adaptive Beamformer als Mikrofon eingesetzt, so verändert sich mit der Richtcharakteristik auch jeweils die Raumimpulsantwort, da etwa Reflexionen ausgeblendet werden. Solche Beamformer besitzen meist eine sehr viel größere Adaptionsgeschwindigkeit und können auch zur Verschlechterung der Adaption des AEC führen.

Da der AEC-Algorithmus besonders zu Beginn seiner Adaption weit von der wahren Impulsantwort

des Echopfads abweicht, wird in der Praxis ein *center-clipper* eingesetzt. Durch seine Nichtlinearität werden kleine Amplituden des Signals abgeschnitten, seine Nichtlinearität kann über die Schätzung der Fehlanpassung (*misalignment*) des AEC gesteuert werden [38].

In der gewünschten Zielanwendung soll der nahe Sprecher zur Förderung der Eigenverständlichkeit auch in den Raum wiedergegeben werden, ähnlich wie der ferne Sprecher. Der Einsatz des AEC kann die Stabilität des Systems gewährleisten und Rückkopplungen unterdrücken, wenn die Adaptionsgenauigkeit stimmt. Ein weiters Problem hier ist aber, dass keine genaue Adaption des AEC mehr möglich ist, da die oben genannten Voraussetzungen für optimale Adaption nicht mehr gegeben sind. Die Adaptionsgeschwindigkeit kann hier auch nicht mehr über eine Pegelwaage gesteuert werden, da sich bei einer solchen Verstärkung das Verhältnis Lautsprecher- zu Mikrofonsignalpegel nicht mehr wesentlich unterscheidet, wenn der ferne oder nahe Sprecher aktiv ist.

Eine Unterdrückung akustischer Echos (*acoustic echo suppression,* AES) [69] ist ein Verfahren, dass nicht mehr die geschätzte Einstreuung des Echopfades phasengenau vom Mikrofonsignal subtrahiert. Eine AES arbeitet im Frequenzbereich und unterdrückt jene Frequenzbänder, die Signale des Echopfades enthalten. Der riesige Vorteil einer AES gegenüber der AEC ist, dass hier weniger Genauigkeit benötigt wird, daher wird der Aufwand kleiner und die Adaptionsgeschwindigkeit größer. Wird die akustischen Echounterdrückung ständig adaptiv betrieben und der nahe Sprecher im Raum verstärkt, werden stationäre Spektralkomponenten der Sprache einfach unterdrückt, so dass im Wesentlichen keine Verstärkung des nahen Sprechers mehr erfolgen kann. Eine Steuerung der Adaption wird dadurch nötig. Es wäre zum Beispiel möglich nur nach fallenden oder steigenden Pegeln eine kurze Zeit die Adaption ein- und auszuschalten. Das bring allerdings eine aufwändige Steuerungslogik mit sich.

5.2.1 Im Zeitbereich

Der Zeitbereichsfilter zur Kompensation akustischer Echos (AEC) ist ein FIR-Filter (*finite impulse response*) mit ausreichender Länge [68]. Die Filterkoeffizienten können über einen NLMS-Adaptionsalgorithmus gefunden werden. Dazu dient die Differenz aus dem gefilterten Lautsprechersignal vom Mikrofonsignal als Adaptionsfehler. Die Kurzzeitenergie des Adaptionsfehlers wird minimal, wenn der Übertragungsweg Lautsprecher-Mirkofon vom Filter richtig identifiziert wird. Das Adaptionsfehlersignal entspricht dann bereits jenem gesuchten Signal,

das von akustischen Echos befreit wurde.

Da ein ideales Anregungssignal weißes Rauschen wäre, um eine schnelle und genaue Identifikation des Echopfades zu ermöglichen, hier aber Sprachsignale transportiert werden, muss ein Dekorrelationsfilter eingesetzt werden. Ein Dekorrelationsfilter soll ein möglichst weißes Signalspektrum erzeugen, das den adaptiven Algorithmen eine schnelle und genaue Adaption ermöglicht. In der Arbeit von [68] wird gezeigt, dass für Sprache bereits eine Nullstelle (FIR Hochpass 1. Ordnung) bei r=0.9 eine wesentliche Verbesserung der Adaptionszeit bringt.

Der Ansatz zur Integration der FIR-Impulsantwort [70] ist zur Reduktion des Berechnungsaufwandes auch erwähnenswert. Dieser Ansatz modelliert nicht die komplette Impulsantwort in der gesamten Bandbreite und erzielt dadurch eine bessere Recheneffizienz. Weil Raumimpulsantworten oft nur spärlich besetzt sind (*sparse*) und für große Verzögerungen weniger hochfrequenten Inhalt besitzen, geht der Ansatz davon aus, dass eine Impulsantwort gefunden werden kann, die spärlich besetzt ist. Diese Impulsantwort soll dann durch Integration zur modellierten Raumimpulsantwort werden. Dabei wird der erste Teil der Impulsantwort genauer modelliert, indem die *sparse* Impulsantwort alle Samples besitzt, und der Nachhall in der Impulsantwort kann durch spärliche Samples in der *sparse* Impulsantwort modelliert werden.

Eine Realisierung mit NLMS benötigt etwa 2M reelle Multiplikationen, 2N+1 Additionen und M Divisionen [42][62].

5.2.2 Im Frequenzbereich

Um die Recheneffizienz der Blockfaltung gegenüber der Faltung im Zeitbereich ausnützen zu können, ist es besonders bei langen Impulsantworten, die eine AEC benötigt, sinnvoll auch die Adaption und Filterung eines Echokompensationsfilters im Frequenzbereich durchzuführen. Ein weiterer Vorteil der Frequenzbereichslösung ist vor allem die schnellere Adaptionsgeschwindigkeit. Der Frequenzbereichsalgrorithmus benötigt nämlich kein weißes Eingangssignal mehr, es ist aber günstig die einzelnen spektralen Adaptionskonstanten in Abhängigkeit der spektralen Signalleistungen zu steuern.

Normalerweise würde eine sehr lange Fouriertransformation des Signals durchgeführt werden, welche sowohl Signal, als auch Impulsantwortlänge fassen kann. Leider ist diese Variante stark latenzbehaftet, weil solche langen FFTs (*fast fourier transformations*) lange Ein- und Ausgangsbuffer benötigen.

Es besteht aber die Möglichkeit die Impulsantwort zu zerteilen und mit entsprechenden Verzögerungen mehrerer adaptiven Frequenzbereichsfiltern zusammenzusetzen.
Ein solcher *fast block LMS* Algorithmus benötigt für einen Signalblock der Länge M etwa 10 M ld (2 M) Multiplikationen. Im Vergleich zum Zeitbereich ergibt sich das Verhältnis [62]:

$$Complexity\ ratio = \frac{5\,ld\,(M)+13}{M}. \tag{107}$$

5.2.3 Im Gammatone-Bereich

Eine akustische Echokompensation im Gammatonebereich macht aus vielerlei Hinsicht Sinn. Einerseits ist die Adaptionsgeschwindigkeit der adaptiven Algorithmen im Frequenzbereich schneller, andererseits der Rechenaufwand geringer.
Um alle Phasenlagen nachbilden zu können, wird eine Gammatone-analyse benötigt, die Real- und Imaginärteil, bzw In-Phase und Quadratur Signale besitzt. Phasenungenauigkeiten bei der Erzeugung des Quadratursignals lassen sich in kleinen Bereichen druch die komplexwertige Filterung kompensieren (Linearkombination). Daher wird zur Bestimmung des Quadratursignals lediglich eine entsprechende Verzögerung um eine Viertelperiode benötigt, welche in etwa eine 90° Phasendrehung erzeugen kann.
Um eine vollständige 10 ms lange Impulsantwort im Gammatone-Bereich mittels komplexwertiger Gewichtung nachzubilden ist es nötig, mehrere Taps zu verwenden.
Die effektive Länge der Kanalimpulsantworten einer Gammatone-Filterbank unterscheidet sich zwischen Filtern hoher und tiefer Mittenfrequenz. Als effektive Länge kann jener Teil angenommen werden, der über der Hälfte der Maximalamplitude der Impulsantwort liegt.
Bei der tiefsten Frequenz ergeben sich bereits beinahe 10 ms effektive Länge, weshalb hier für die Filterung lediglich 2 komplexe Gewichte benötigt werden um eine 10 ms lange Impulsantwort zu modellieren.
Bei der höchsten Frequenz ergibt sich etwa 0.2 ms effektive Länge. Um hier eine Impulsanwort von ganzen 10 ms nachbilden zu können, müssen 50 Verzögerungsglieder verwendet werden. An jedem Verzögerungsglied erfolgt die Bestimmung des Quadratursignals und eine komplexwertige Gewichtung.

6 Literaturverzeichnis

[1] J.C. Junqua, "*The Lombard Reflex and its Role on Human Listeners and Automatic Speech Recognizers*", Journal of the Acoustical Society of America, Vol. 93, No. 1 (510-524), ASA, , 1993.

[2] Cornelia Falch, "*Smart Microphone Array, Part II*", Institute of Electronic Music and Acoustics, Graz in cooperation with AKG acoustics GmbH, Vienna, 2003, Unveröffentlicht.

[3] Patent: International application No. PCT/SG99/00119, International Publication Number. WO 00/30264, Invention of Siew K. Hui, 10 Science Park Road #03-20, The Alpha, Singapore Science Park II, Singapore 117684 (SG), Austrian Patent Office, Kohlmarkt 8-10, A-1014 Vienna, 25. Mai 2000.

[4] Xianxian Zhang, John H. L. Hansen, "*CSA-BF: A Constrained Switched Adaptive Beamformer for Speech Enhancement and Recognition in Real Car Environments*", IEEE Transactions on Speech and Audio Processing, Vol. 11, No. 6 (733-745), IEEE, Nov, 2003.

[5] Radu Balan, Justinian Rosca, "*Microphone Array Speech Enhancement by Bayesian Estimation of Spectral Amplitude and Phase*", IEEE Sensor Array and Multichannel Signal Processing Workshop Proceedings, August, 2002.

[6] Martin Pflüger, "*Modelle des peripheren Gehörs am Beispiel der menschlichen Lautheitsempfindung*", Dissertation, Institut für Elektronische Musik und Akustik, Musikhochschule Graz, 1997.

[7] M. Pflueger, R. Hoeldrich, "*Nonlinear All-Pole and One-Zero Gammatone Filters*", acta acustica, Vol. 83, (513-519), Hirzel Verlag, , 1997.

[8] E. Zwicker, H. Fastl, "*Psychoacoustics, facts and models*", Springer, Berlin Heidelberg, 1999.

[9] E. Zwicker, R. Feldtkeller, "*Das Ohr als Nachrichtenempfänger*", Hirzel Verlag Stuttgart, 1967.

[10] Ernst Terhardt, "*Akustische Kommunikation*", Springer, Berlin Heidelberg, 1998.

[11] L. Lin, E. Ambikairajah, W. H. Holmes, "*Auditory Filterbank Design Using Masking Curves*", Proc. EUROSPEECH Scandinavia, 7th European Conference on Speech Communication and Technology, 2001.

[12] L. Lin, E. Ambikairajah, W. H. Holmes, "*Auditory Filter Bank Inversion*", The 2001 IEEE International Symposium, ISCAS, IEEE Circuits and Systems, (537-540), 6-9 May, 2001.

[13] L. Lin, E. Ambikairajah, W. H. Holmes, "*Perceptual Domain Based Speech and Audio Coder*", Proc. of the third international symposium DSPCS 2002, Sydney, January 28-31, 2002.

[14] R. F. Lyon, C. A. Mead, "*Cochlear Hydrodynamics Demystified*", Technical Report, , California Institute of Technology, 1988.

[15] R. F. Lyon, "*The All-Pole Gammatone Filter and Auditory Models*", Proc. Forum Acusticum,

Antwerpen, 1996.

[16] G. Kubin, W. B. Kleijn, "*On Speech Coding in a Perceptual Domain*", Proc. ICASSP Phoenix USA, International Conference on Acoustics, Speech, and Signal Processing, (205-208), 15-19 March, 1999.

[17] Malcolm Slaney, "*An Efficient Implementation of the Patterson-Holdsworth Auditory Filter Bank*", Technical Report #35, Apple Computer, Inc., Perception Group – Advanced Technical Group, 1993.

[18] Malcolm Slaney, Daniel Naar, Richard F. Lyon, "*Auditory Model Inversion for Sound Separation*", , IEEE International Conference on Acoustics, Speech, and Signal Processing, (77-80), 19-22 April, 1994.

[19] Gerhard Stoll, John G. Beerends, Roland Bitto, Karlheinz Brandenburg, Catherine Colomes, Bernhard Feiten, Michael Keyhl, Christian Schmidmer, Thomas Sporer, Thilo Thiede, William C. Treurniet, "*PEAQ - der neue ITU-Standard zur objektiven Messung der wahrgenommenen Audioqualität*", RTM - Rundfunktechnische Mitteilungen, die Fachzeitschrift für Hörfunk und Fernsehtechnik, 43. Jahrgang, ISSN 0035-9890 (81-120), Firma Mensing GmbH + Co. KG, Abteilung Verlag, Sept, 1999.

[20] Frank Baumgarte, "*Ein psychophysiologisches Gehörmodell zur Nachbildung von Wahrnehmungsschwellen für die Audiocodierung*", Dissertation, Institut für Theoretische Nachrichtentechnik und Informationsverarbeitung, Universität Hannover, 2000.

[21] T. Irino, R. D. Patterson, "*A time-domain, level-dependent auditory filter: The gammachirp*", Journal of the Acoustical Society of America, Vol. 101, No. 1 (412-419), ASA, Jan, 1997.

[22] T.Irino, M. Unoki, "*An Analysis/Synthesis Auditory Filterbank Based on an IIR Implementation of the* ", Journal of the Acoustical Society of Japan, Vol. 20, No. 5 (397-406), ASJ, Nov, 1999.

[23] Toshio Irino, "*Noise Suppression Using a Time-Varying Analysis/Synthesis Gammachirp Filterbank*", Proc. ICASSP Phoenix USA, International Conference on Acoustics, Speech, and Signal Processing, (97-100), 15-19 March, 1999.

[24] Joachim Thiemann, "*Acoustic Noise Suppression for Speech Signals using Auditory Masking Effects*", Dissertation, Department of Electrical & Computing Engineering, McGill University, Montreal, Canada, 2001.

[25] D. Ellis, D. Rosenthal, "*Mid-Level representations for computational auditory scene analysis: the weft element*", , D.F. Rosenthal, H. Okuno, Mahwah, 1998.

[26] A. S. Bregman, "*Auditory Scene Analysis, The Perceptual Organization of Sounds*", Cambridge Mass., MIT Press, 1990.

[27] E. D. Scheirer, "*Sound Scene Segmentation by Dynamic Detection of Correlogram Comodulation*", Technical Report #491, Media Laboratory Perceptual Computing, MIT, 1999.

[28] M. Unoki, M. Agaki, "*A Method of Signal Extraction from Noisy Signal Based on Auditory Scene Analysis*", Speech Communication, Vol. 27, No. 3-4 (261-279), Elsevier, April, 1999.

[29] D. L. Wang, G. J. Brown, "*Separation of Speech from Interfering Sounds Based on Oscillatory Correlation*", IEEE Transactions on Neural Networks, Vol. 10, No. 3 (684-697), IEEE, May, 1999.

[30] Yariv Ephraim, David Malah, "*Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator*", IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. ASSP-32, No. 6 (1109-1121), IEEE, Dec, 1984.

[31] Yariv Ephraim, David Malah, "*Speech Enhancement Using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator*", IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. ASSP-33, No. 2 (443-445), IEEE, April, 1985.

[32] Israel Cohen, "*Noise Spectrum Estimation in Adverse Environments: Improved Minima Controlled Recursive Averaging*", IEEE Transactions on Speech and Audio Processing, Vol. 11, No. 5 (466-475), IEEE, Sept, 2003.

[33] I. Cohen, B. Berdugo, "*Speech Enhancement for Non-Stationary Noise Environments*", Signal Processing, , No. 11 (2403-2418), Elsevier, Nov, 2001.

[34] Israel Cohen, "*On The Decision-Directed Estimation Approach of Ephraim and Malah*", Proc. ICASSP 04, International Conference on Acoustics, Speech, and Signal Processing, (293-296), 17-21 May, 2004.

[35] Israel Cohen, "*Speech Enhancement Using a Noncausal A Priori SNR Estimator*", Signal Processing Letters, , No. 9 (725-728), IEEE, Sept, 2004.

[36] Olivier Cappé, "*Elimination of the Musical Noise Phenomenon with the Ephraim and Malah Noise Suppressor*", IEEE Transactions on Speech and Audio Processing, Vol. 2, No. 2 (345-349), IEEE, April, 1994.

[37] Yi Hu, Philipos C. Loizou, "*Incorporating a Psychoacoustical Model in Frequency Domain Speech Enhancement*", IEEE Signal Processing Letters, Vol. 11, No. 2 (270-273), IEEE, Feb, 2004.

[38] P. Vary, U. Heute, W. Hess, "*Digitale Sprachsignalverarbeitung*", Teubner Stuttgart, 1998.

[39] D. E. Tsoukalas, J. N. Mourjopoulos, G. Kokkinakis, "*Speech Enhancement Based on Audible Noise Suppression*", IEEE Transactions on Speech and Audio Processing, Vol. 5, No. 6 (497-514), IEEE, Nov, 1997.

[40] Philipp M. Krejci, "*Psychoakustik*", , , 2000.

[41] D. L. Wang, G. J. Brown, "*Separation of Speech from Interfering Sounds Based on Oscillatory Correlation*", IEEE Transactions on Neural Networks, Vol. 10, No. 3 (684-697), IEEE, May, 1999.

[42] A. V. Oppenheim, R. W. Schafer, J. R. Buck, "*Discrete-Time Signal Processing*", Prentice Hall, 1999.

[43] P. Dutilleux, U. Zölzer, "*DAFX*", Wiley&Sons LTD, 2002.

[44] T. Irino, R. D. Patterson, "*A Compressive Gammachirp Auditory Filter for Both Physiological and Psychophysical Data*", Journal of the Acoustical Society of America, Vol. 109, No. 5 (2008-2022), ASA, May, 2001.

[45] Rainer Martin, "*Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics*", IEEE Transactions on Speech and Audio Processing, Vol. 9, No. 5 (504-512), IEEE, July, 2001.

[46] Wolfgang Hess, "*A Pitch-Synchronous Digital Feature Extraction System for Phonemic Recognition of Speech*", IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. ASSP-24, No. 1 (14-25), IEEE, Feb, 1976.

[47] Gerhard Doblinger, "*Computational Efficient Speech Enhancement by Spectral Minima Tracking in Subbands*", Proc. EUROSPEECH Madrid, European Conference on Speech Technology and Communication, Seiten 1513-1516, 1995.

[48] F. Jabloun, A. E. Çetin, Engin Erzin, "*Teager Energy Based Feature Parameters for Speech Recognition in Car Noise*", IEEE Signal Processing Letters, Vol. 6, No. 10 (259-261), IEEE, Oct, 1999.

[49] J. Meyer, K. Simmer, K. Kammeyer, "*Comparison of One- and Two-Channel Noise-Estimation Techniques*", Proc. 5th IWAENC 97, London, Seiten 137-145, 11-12 Sept, 1997.

[50] L. Arslan, A. McCree, V. Viswanathan, "*New Methods for Adaptive Noise Suppression*", Proc. ICASSP 95, International Conference on Acoustics, Speech, and Signal Processing, (9-12), 9-12 May, 1995.

[51] M. Berouti, R. Schwartz, J. Makhoul, "*Enhancement of Speech Corrupted by Acoustic Noise*", Proc. ICASSP 79, International Conference on Acoustics, Speech, and Signal Processing, (208-211), April, 1979.

[52] H. G. Hirsch, C. Ehrlicher, "*Noise Estimation Techniques for Robust Speech Recognition*", Proc. ICASSP 95, International Conference on Acoustics, Speech, and Signal Processing, (153-156), 9-12 May, 1995.

[53] V. Stahl, A. Fischer, R. Bippus, "*Quantile Based Noise Estimation for Spectral Subtraction and Wiener Filtering*", Proc. ICASSP 00, International Conference on Acoustics, Speech, and Signal Processing, (1875-1878), 5-9 June, 2000.

[54] D. Ealey, H. Kelleher, D. Pearce, "*Harmonic Tunneling: Tracking Non-Stationary Noises During Speech*", Proc. EUROSPEECH, 2001.

[55] N. W. D. Evans, J. S. Mason, "*Time-Frequency Quantile-Based Noise Estimation*", EUSIPCO 02,

2002.

[56] Nathalie Virag, "*Single Channel Speech Enhancement Based on Masking Properties of the Human Auditory System*", IEEE Transactions on Speech and Audio Processing, Vol. 7, No. 2 (126-137), IEEE, March, 1999.

[57] J. Thiemann, P. Kabal, "*Low Distortion Acoustic Noise Suppresseion Using a Perceptual Model for Speech Signals*", IEEE Workshop Proceedings, Workshop on Speech Coding, (172-174), 6-9. Oct, 2002.

[58] S. Gustafsson, P. Jax, P. Vary, "*A Novel Psychoacoustically Motivated Audio Enhancement Algorithm Preserving Background Noise Characteristics*", Proc. ICASSP 98, International Conference on Acoustics, Speech, and Signal Processing, (397-400), May, 1998.

[59] L. Lin, W. H. Holmes, E. Ambikairajah, "*Subband Noise Estimation for Speech Enhancement Using a Perceptual Wiener Filter*", Proc. ICASSP 03, International Conference on Acoustics, Speech, and Signal Processing, (80-83), 6-10 April, 2003.

[60] Robert Höldrich, Markus Lorber, "*Non-Linear Spectral Subtraction with Combined Smoothing Strategies for Broadband Noise Reduction*", 103rd AES-Convention 97, New York, 26-29 September, 1997.

[61] I.N. Bronstein, K. A. Semandjajew, G. Musiol, H. Mühlig, "*Taschenbuch der Mathematik*", Verlag Harri Deutsch, Thun und Frankfurt am Main, 2001.

[62] Simon Haykin, "*Adaptive Filter Theory*", Prentice Hall, 2002.

[63] M. T. Johnson, A. C. Lindgren, R. J. Povinelli, X. Yuan, "*Performance of Nonlinear Speech Enhancement Using Phase Space Reconstruction*", Proceedings of the ICASSP 03, IEEE International Conference on Acoustics, Speech, and Signal Processing, (920-923), 6-10 April, 2003.

[64] P. J. Wolfe, S. J. Godsill, "*Simple Alternatives to the Ephraim and Malah Suppression Rule for Speech Enhancement*", Proc. IEEE Signal Processing Workshop , 11th IEEE Signal Processing Workshop, (496-499), 6-8 Aug, 2001.

[65] R. Martin, I. Wittke, P. Jax, "*Optimized Estimation of Spectral Parameters for the Coding of Noisy Speech*", Proc. ICASSP 00, International Conference on Acoustics, Speech, and Signal Processing, (1479-1482), 5-9 June, 2000.

[66] D. Malah, R. V. Cox, A. J. Accardi, "*Tracking Speech-Presence Uncertainty to Improve Speech Enhancement in Non-Stationary Noise Environments*", Proc. ICASSP 99, International Conference on Acoustics, Speech, and Signal Processing, (789-792), 15-19 March, 1999.

[67] I. Cohen, B. Berdugo, "*Speech Enhancement Based on a Microphone Array and Log-Spectral Amplitude Estimation*", The 22nd Convention of Electrical and Electronics Engineers in Israel, Seiten

4-6, 1 Dec., 2002.

[68] C. Breining, P. Dreiscitel, E. Hansler, A. Mader, B. Nitsch, H. Puder, T. Schertler, G. Schmidt, J. Tilp, "*Acoustic Echo Control. An Applocation of Very-High-Order Adaptive Filters*", Signal Processing Magazine, Vol. 16, No. 4 (42-69), IEEE, July, 1999.

[69] Fredrik Wallin, "*Combining Acoustic Echo Cancellation and Suppression*", Dissertation, Division of Automatic Control and Communication Systems, Department of Electrical Engineering, Linköping University, Sweden, 2003.

[70] R. L. M. Heylen, Malcolm O. Hawksford, "*The Integrating Finite Impulse Response Filter*", The 94th Convention of the Audio Engineering Society, Berlin, 16-19 March, 1993.

7 Anhang: Berechnungen zu den Gammatone-Filtern

Contents

1. Gammatone

The gammatone function models the cochlea motion and can be used to describe the frequency analysis of the human inner ear. For further reference the paper of Lyon [1] on all-pole auditory models and the gammatone is recommended.

2. Continuous-Time Description

The gammatone function $g_m(t)$ of the order m is given by a harmonic oscillation modulated with the gamma-function Γ.

$$g_m(t) = t^{m-1} e^{-bt} \cos(\omega t + \phi) \tag{2.1}$$

m	... gammatone order
$b = 2\pi\Delta\omega$	... bandwidth
$w = 2\pi f$	... center/carrier frequency
ϕ	... carrier phase angle

(2.2)

Usually the phase-term ϕ is set to zero, leading to the cos-phase gammatone.

3. Laplace-Domain Description

In order to obtain the gammatone's Laplace-Domain description, we have to transform the function given in equation 2.1 into the Laplace-Domain.

Following Laplace transform pairs and properties are helpful for this purpose:

$$e^{at} \cdot u(t) \quad \overset{L}{\longleftrightarrow} \quad \frac{1}{s-a} \tag{3.1}$$

$$e^{j\omega t} x(t) \cdot u(t) \quad \overset{L}{\longleftrightarrow} \quad X(s-j\omega) \tag{3.2}$$

$$t^m x(t) \cdot u(t) \quad \overset{L}{\longleftrightarrow} \quad (-1)^m \frac{d^m}{ds^m} X(s) \tag{3.3}$$

A 1^{st} order gammatone envelope in the Laplace domain looks like a simple low-pass, using the exponential function transform pair we obtain

$$G_{env,1}(s) = \frac{1}{s+b}. \tag{3.4}$$

Arbitrarily, the m^{th} order gammatone envelope can be derived from $G_{env,1}(s)$ as an arbitrary expression by applying the derivation theorem

$$G_{env,m}(s) = (-1)^{m-1} \frac{(m-1)!}{(s+b)^m}. \tag{3.5}$$

Using the modulation theorem from above the complex $e^{j\omega t}$-modulated (analytic) gammatone is

$$G_{analytic,m}(s) = (-1)^{m-1} \frac{(m-1)!}{(s+b-j\omega)^m}. \tag{3.6}$$

To obtain a real-valued cos-phase gammatone we can decompose the cos-modulation from complex modulations, using the property

$$\cos(\alpha) = \frac{1}{2} \left(e^{j\alpha} + e^{-j\alpha} \right). \tag{3.7}$$

Now the whole description of the m^{th} order gammatone can be expressed by one equation

$$\begin{aligned} G_m(s) &= \frac{(-1)^{m-1}(m-1)!}{2} \left[\frac{1}{(s+b-j\omega)^m} + \frac{1}{(s+b+j\omega)^m} \right] \\ &= \frac{(-1)^{m-1}(m-1)!}{2} \frac{(s+b-j\omega)^m + (s+b+j\omega)^m}{\left[(s+b)^2+\omega^2)\right]^m}. \end{aligned} \tag{3.8}$$

4. Discrete-Time Description

A simple discrete-time description of the gammatone can be obtained by using the impulse invariance technique, i.e. a sampled version of the continuous-time gammatone function 2.1, by substituting $t = n \cdot T$, with T being the sampling interval. (This is only allowed if the gammatone is band-limited, you can find this property in Slaney's work on the gammatone [2].)

We assume that the gammatone is bandlimited and get following expression

$$g_m[n] = T^{m-1} n^{m-1} e^{-Bn} \cos(\theta n + \phi), \tag{4.1}$$

wherein $B = b \cdot T$ and the normalized frequency $\theta = \omega T$.

5. z-Domain Description

There are three ways to obtain the z-domain description of the gammatone we want to discuss here.

- impulse invariance
  - z-transform of the discrete-time gammatone (eq. 4.1)
  - pole transform from continuous-time Laplace-domain to the discrete-time z-domain (eq. 3.8)
- bilinear transform

5.1. **z-transform of the discrete-time gammatone.** We want to use the z-transform on the sampled gammatone given by equation 4.1.

Following z-Transform pairs and properties are helpful for this purpose:

$$a^n \cdot u[n] \quad \overset{z}{\longleftrightarrow} \quad \frac{1}{1-az^{-1}} \tag{5.1}$$

$$b^n x[n] \cdot u[n] \quad \overset{z}{\longleftrightarrow} \quad X\left(z/b\right) \tag{5.2}$$

$$nx[n] \cdot u[n] \quad \overset{z}{\longleftrightarrow} \quad -z^{-1}\frac{d}{dz}X(z) \tag{5.3}$$

Applying the z-transform to the unmodulated 1^{st} order gammatone envelope, and substituting $r = e^{-B}$ we simply get

$$G_{env,1}(z) = \frac{1}{1-rz^{-1}}. \tag{5.4}$$

Higher order gammatone envelope functions ($m > 1$) can be obtained by applying the derivation theorem

$$G_{env,m}(z) = -T^{m-1}z^{-1}\frac{d}{dz}G_{env,m-1}(z). \tag{5.5}$$

Unfortunately this is no closed expression for all m, like equation 3.8 in the Laplace-domain, thus we can only use this equation to derive the gammatone descriptions step-by-step.

We only consider the first order gammatone here, because it turns out that starting in the Laplace-domain is much easier. The analytic gammatone is obtained by using the modulation theorem with $e^{j\theta n}$

$$G_{analytic,1}(z) \quad = \quad \frac{1}{1-re^{j\theta}z^{-1}}. \tag{5.6}$$

5.1.1. 1^{st} *order gammatone.* The cos-phase gammatone is found to be (using eq. 3.7)

$$\begin{aligned} G_1(z) &= \frac{1}{2}\left[\frac{1}{1-re^{j\theta}z^{-1}} + \frac{1}{1-re^{-j\theta}z^{-1}}\right] \\ &= \frac{1-r\cos(\theta)z^{-1}}{1-2r\cos(\theta)z^{-1}+r^2z^{-2}}. \end{aligned} \tag{5.7}$$

From this point we want to skip to the next subsection for the following reasons:

- up to this point no closed expression for the m^{th} order gammatone's z-domain description was found. We would have to derive it step-by-step, the form of the numerator polynome is hard to predict.
- applying real-valued modulation changes the numerator polynome so that we don't know which zeros we get. It turns out that a factorization of the numerator polynomial is quite hard for higher order gammatones

5.2. **pole transform from continuous-time Laplace-domain to the discrete-time z-domain.** Impulse variance technique can also be done by converting each complex Laplace-domain pole into a z-domain pole. Thus we have to use the partial fraction expansion of the Laplace-domain function we want to convert. If a cascade form of real-valued rational functions functions exists in the Laplace domain, it's also allowed to convert each of the cascade's sections separately into the z-domain, subsequently putting all parts together again. This is how Slaney derived his 4^{th} order gammatone implementation [2].

Assuming a single cascade stage $F(s)$ with one zero s_z and a complex conjugate pole-pair $s_p = -b \pm +j\omega$ in its Laplace-domain description, we get

$$\begin{aligned} F(s) &= \frac{s - s_z}{(s+b)^2 + \omega^2} \\ &= \frac{s_z + b - j\omega}{j2\omega} \frac{1}{s + b + j\omega} - \frac{s_z + b - j\omega}{j2\omega} \frac{1}{s + b - j\omega}. \end{aligned} \tag{5.8}$$

Converting each pole $s_p = -b \pm j\omega$ to its z-domain description $z_p = re^{\pm j\theta}$ and multiplying with T, ($r = e^{-B}$, $B = b \cdot T$, and $\theta = \omega \cdot T$), we obtain the z-domain cascade stage $F(z)$

$$\begin{aligned} F(z) &= T \cdot \frac{s_z + b - j\omega}{j2\omega} \frac{1}{1 - re^{-j\theta}z^{-1}} - T \cdot \frac{s_z + b - j\omega}{j2\omega} \frac{1}{1 - re^{+j\theta}z^{-1}} \\ &= T \cdot \frac{1 - rz^{-1}\left[\frac{(s_z+b)}{\omega}\sin(\theta) + \cos(\theta)\right]}{1 - 2r\cos(\theta)z^{-1} + r^2z^{-2}}. \end{aligned} \tag{5.9}$$

Now we can derive a few gammatone functions from their Laplace-domain description:

5.2.1. 2^{nd} *order gammatone.* The Laplace-domain description of the 2^{nd} order gammatone is

$$\begin{aligned} G_2(s) &= -\frac{1}{2}\left[\frac{1}{(s + b - j\omega)^2} + \frac{1}{(s + b + j\omega)^2}\right] \\ &= -\frac{(s+b)^3 - 3\omega^2(s+b)}{((s+b)^2 + \omega^2)^2}. \end{aligned} \tag{5.10}$$

The zeros of the numerator are found to be (I used Mapple)

$$s_{z,i} = -b + \omega, -b - \omega. \tag{5.11}$$

We can see that the Laplace-domain function can be expressed in a cascade form:

$$G_2(s) = -T^2 \cdot \frac{s + b + \omega}{(s+b)^2 + \omega^2} \frac{s + b - \omega}{(s+b)^2 + \omega^2}. \tag{5.12}$$

Using now the above equations 5.8 and 5.9, we can immediately write the z-transform of the 2^{nd} order gammatone

$$\begin{aligned} G_2(z) &= -T^2 \cdot \frac{1 - rz^{-1}(\sin(\theta) + \cos(\theta))}{1 - 2r\cos(\theta)z^{-1} + r^2z^{-2}} \cdot \frac{1 - rz^{-1}(-\sin(\theta) + \cos(\theta))}{1 - 2r\cos(\theta)z^{-1} + r^2z^{-2}} \\ &= -T^2 \cdot \frac{1 - r\sqrt{2} \cdot z^{-1}\cos\left(\theta - \frac{\pi}{4}\right)}{1 - 2r\cos(\theta)z^{-1} + r^2z^{-2}} \cdot \frac{1 - r\sqrt{2} \cdot z^{-1}\cos\left(\theta + \frac{\pi}{4}\right)}{1 - 2r\cos(\theta)z^{-1} + r^2z^{-2}}. \end{aligned} \tag{5.13}$$

5.2.2. 3^{rd} *order gammatone.* The 3^{rd} order gammatone expression in the Laplace-domain is

$$\begin{aligned} G_3(s) &= \frac{(3-1)!(-1)^{3-1}}{2}\left[\frac{1}{(s+b-j\omega)^3}+\frac{1}{(s+b+j\omega)^3}\right] \\ &= 2\cdot\frac{(s+b)^3-3\omega^2(s+b)}{((s+b)^2+\omega^2)^3} \\ &= 2\cdot\frac{(s+b)(s+b+\sqrt{3}\omega)(s+b-\sqrt{3}\omega)}{((s+b)^2+\omega^2)^3}, \end{aligned} \tag{5.14}$$

with the zeros

$$s_{z,i} \quad = -b, -b+\sqrt{3}\omega, -b-\sqrt{3}\omega. \tag{5.15}$$

We have found the 3^{rd} order gammatone's z-domain description as

$$\begin{aligned} G_3(z) &= 2\cdot T^3\cdot\frac{1-rz^{-1}\cos(\theta)}{1-2r\cos(\theta)z^{-1}+r^2z^{-2}}\cdot\frac{1-rz^{-1}\left(\sqrt{3}\sin(\theta)+\cos(\theta)\right)}{1-2r\cos(\theta)z^{-1}+r^2z^{-2}}\cdot \\ &\quad \frac{1-rz^{-1}\left(-\sqrt{3}\sin(\theta)+\cos(\theta)\right)}{1-2r\cos(\theta)z^{-1}+r^2z^{-2}} \\ &= 2\cdot T^3\cdot\frac{1-rz^{-1}\cos(\theta)}{1-2r\cos(\theta)z^{-1}+r^2z^{-2}}\cdot\frac{1-\frac{r}{2}\cdot z^{-1}\cos\left(\theta-\frac{\pi}{3}\right)}{1-2r\cos(\theta)z^{-1}+r^2z^{-2}}\cdot \\ &\quad \frac{1-\frac{r}{2}\cdot z^{-1}\cos\left(\theta+\frac{\pi}{3}\right)}{1-2r\cos(\theta)z^{-1}+r^2z^{-2}}. \end{aligned} \tag{5.16}$$

5.2.3. 4^{th} *order gammatone.* The 4^{th} order gammatone in the Laplace-domain is

$$\begin{aligned} G_4(s) &= \frac{3!(-1)^3}{2}\left[\frac{1}{(s+b-j\omega)^4}+\frac{1}{(s+b+j\omega)^4}\right] \\ &= -6\cdot\frac{s^4+4bs^3+6(b^2-\omega^2)+4(b^3-3b\omega^2)s+b^4+6b^2\omega^2+\omega^4}{((s+b)^2+\omega^2)^4} \end{aligned} \tag{5.17}$$

Similar to Slaney's Mathematica solution [2] we get here

$$s_{z,i} = -b\pm(1\pm\sqrt{2})\omega. \tag{5.18}$$

Writing the equations into a cascaded form

$$\begin{aligned} G_4(s) &= -6\cdot T^4\cdot\frac{s+b+(1+\sqrt{2})\omega}{(s+b)^2+\omega^2}\cdot\frac{s+b+(1-\sqrt{2})\omega}{(s+b)^2+\omega^2}\cdot \\ &\quad \frac{s+b-(1-\sqrt{2})\omega}{(s+b)^2+\omega^2}\cdot\frac{s+b-(1+\sqrt{2})\omega}{(s+b)^2+\omega^2}. \end{aligned} \tag{5.19}$$

$$
\begin{aligned}
G_4(z) &= -6 \cdot T^4 \cdot \frac{1 - rz^{-1}\left[\left(1+\sqrt{2}\right)\sin(\theta) + \cos(\theta)\right]}{1 - 2r\cos(\theta)z^{-1} + r^2 z^{-2}} \cdot \\
&\quad \frac{1 - rz^{-1}\left[\left(1-\sqrt{2}\right)\sin(\theta) + \cos(\theta)\right]}{1 - 2r\cos(\theta)z^{-1} + r^2 z^{-2}} \cdot \\
&\quad \frac{1 - rz^{-1}\left[-\left(1-\sqrt{2}\right)\sin(\theta) + \cos(\theta)\right]}{1 - 2r\cos(\theta)z^{-1} + r^2 z^{-2}} \cdot \\
&\quad \frac{1 - rz^{-1}\left[-\left(1+\sqrt{2}\right)\sin(\theta) + \cos(\theta)\right]}{1 - 2r\cos(\theta)z^{-1} + r^2 z^{-2}} \\
&= -6 \cdot T^4 \cdot \frac{1 - r\sqrt{4+2\sqrt{2}} \cdot z^{-1}\cos\left(\theta - \frac{3\pi}{8}\right)}{1 - 2r\cos(\theta)z^{-1} + r^2 z^{-2}} \cdot \\
&\quad \frac{1 - r\sqrt{4-2\sqrt{2}} \cdot z^{-1}\cos\left(\theta - \frac{\pi}{8}\right)}{1 - 2r\cos(\theta)z^{-1} + r^2 z^{-2}} \cdot \\
&\quad \frac{1 - r\sqrt{4-2\sqrt{2}} \cdot z^{-1}\cos\left(\theta + \frac{\pi}{8}\right)}{1 - 2r\cos(\theta)z^{-1} + r^2 z^{-2}} \cdot \\
&\quad \frac{1 - r\sqrt{4+2\sqrt{2}} \cdot z^{-1}\cos\left(\theta + \frac{3\pi}{8}\right)}{1 - 2r\cos(\theta)z^{-1} + r^2 z^{-2}}.
\end{aligned} \tag{5.20}
$$

From these z-domain descriptions we can see that the cos-phase m^{th} order gammatone has m identical complex conjugate pole-pairs at the positions $re^{\pm j\theta}$, and m zeros on the real axis, the positions of which depend on the frequency and bandwidth coefficient θ and r, respectively.

5.3. **sin-phase gammatones.** If we used two gammatone filters, one with a sine carrier and another one with a cosine carrier, in parallel, we would elegantly get the squared amplitude envelope suming up both squared output signals. Of course two filters in parallel wouldn't be very efficient, but there are a few hints that indicate a possible low effort computation of those parallel filters.

If the carrier wave in the gammatone equation is changed from a cosine to a sine (equ 2.1) the Laplace-domain description changes. Instead of m zeros, only $m-1$ zeros remain, with m being the gammatone order (see also Slaney [2]). Now how does the carrier phase affect the z-Transform of the gammatone function?

Looking at the 3^{rd} order Gammatone we get

$$
G_3(s) = 2 \cdot \prod_{i=1}^{3} \frac{(s - s_{z,i})}{((s+b)^2 + \omega^2)}, \tag{5.21}
$$

with the Laplace-domain zeros

$$
s_{z,i} = -b, -b + \sqrt{3}\omega, -b - \sqrt{3}\omega. \tag{5.22}
$$

Using a sin-phase carrier, the Laplace-domain zeros are

$$
s_{z,i} = -b + \frac{\sqrt{3}}{3}\omega, -b - \frac{\sqrt{3}}{3}\omega. \tag{5.23}
$$

We do not really see a special relationship in the resulting Laplace-domain comparison, thus we want to look at the z-transform pairs. Therefore we need to find

a new mapping of a Laplace-domain cascade stage with a complex conjugate pole and no zeros at first. This occurs similar to equation 5.8:

$$\begin{aligned} F_2(s) &= \frac{1}{(s+b)^2+\omega^2} \\ &= \frac{1}{j2\omega}\frac{1}{s+b+j\omega} - \frac{1}{j2\omega}\frac{1}{s+b-j\omega}. \end{aligned} \tag{5.24}$$

Converting again each pole $s_p = -b \pm j\omega$ to its z-domain description $z_p = re^{\pm j\theta}$ and multiplying with T, ($r = e^{-B}$, $B = b \cdot T$, and $\theta = \omega \cdot T$), we obtain the z-domain cascade stage $F_2(z)$

$$\begin{aligned} F_2(z) &= \frac{T}{\omega} \cdot \frac{1}{2j}\left[\frac{1}{1-re^{-j\theta}z^{-1}} - T \cdot \frac{1}{j2\omega}\frac{1}{1-re^{+j\theta}z^{-1}}\right] \\ &= T \cdot \frac{r\sin(\theta)z^{-1}}{1-2r\cos(\theta)z^{-1}+r^2z^{-2}}. \end{aligned} \tag{5.25}$$

So the sin-phase gammatone function turns out to be

$$\begin{aligned} G_{3,\sin}(z) &= 2 \cdot T^3 \cdot \frac{r\sin(\theta)z^{-1}}{1-2r\cos(\theta)z^{-1}+r^2z^{-2}} \cdot \frac{1-rz^{-1}\left(\frac{\sqrt{3}}{3}\sin(\theta)+\cos(\theta)\right)}{1-2r\cos(\theta)z^{-1}+r^2z^{-2}} \cdot \\ &\quad \frac{1-rz^{-1}\left(-\frac{\sqrt{3}}{3}\sin(\theta)+\cos(\theta)\right)}{1-2r\cos(\theta)z^{-1}+r^2z^{-2}} \\ &= 2 \cdot T^3 \cdot \frac{r\sin(\theta)z^{-1}}{1-2r\cos(\theta)z^{-1}+r^2z^{-2}} \cdot \frac{1-r\frac{2}{\sqrt{3}}\cdot z^{-1}\cos\left(\theta-\frac{\pi}{6}\right)}{1-2r\cos(\theta)z^{-1}+r^2z^{-2}} \cdot \\ &\quad \frac{1-r\frac{2}{\sqrt{3}}\cdot z^{-1}\cos\left(\theta+\frac{\pi}{6}\right)}{1-2r\cos(\theta)z^{-1}+r^2z^{-2}}. \end{aligned} \tag{5.26}$$

A way to turn the 3^{rd} order sin-phase gammatone into a cos-phase gammatone, we could use a phase shifter build out of the relation of both z-transforms

$$\begin{aligned} H_{3,90^\circ}(z) &= \frac{\left(1-rz^{-1}\cos(\theta)\right)\left(1-\frac{r}{2}\cdot z^{-1}\cos\left(\theta-\frac{\pi}{3}\right)\right)}{r\sin(\theta)z^{-1}\left(1-r\frac{2}{\sqrt{3}}\cdot z^{-1}\cos\left(\theta-\frac{\pi}{6}\right)\right)} \cdot \\ &\quad \frac{\left(1-\frac{r}{2}\cdot z^{-1}\cos\left(\theta+\frac{\pi}{3}\right)\right)}{\left(1-r\frac{2}{\sqrt{3}}\cdot z^{-1}\cos\left(\theta+\frac{\pi}{6}\right)\right)}. \end{aligned} \tag{5.27}$$

It would be quite a simplification if this phase shifter could be applied to an arbitrary gammatone. Another drawback is that in case of the 2^{nd} and 4^{th} order gammatone, such a filter wouldn't be stable. Generously cancelling some zeros with nearby poles leads to the rough approximation

$$\tilde{H}_{m,90^\circ}(z) = \frac{1-rz^{-1}\cos(\theta)}{r\sin(\theta)z^{-1}} = \frac{z}{\sin(\theta)} - \cot(\theta). \tag{5.28}$$

This system is not causal, so we apply a delay to the whole transfer function

$$\tilde{H}_{m,90^\circ}(z) \simeq \frac{1}{\sin(\theta)} - \cot(\theta)z^{-1}. \tag{5.29}$$

Unfortunately the resulting system doesn't work well. The system's zero lies at $z = \cos(\theta)$, thus for $\theta > 0$ it is a minimum phase system. This somehow explains the bad performance.

We will make another attempt to derive a simple 90-phase shifter using the 1^{st} order APGF's carrier (see also section 7).

$$H_{1,\mathrm{APGF}}(z) = \frac{1}{1 - 2r\cos(\theta)z^{-1} + r^2z^2} \quad \overset{z^{-1}}{\longleftrightarrow} \quad h_{1,\mathrm{APGF}}[n] = r^n \frac{\sin[\theta(n+1)]}{\sin(\theta)} u[n]. \tag{5.30}$$

Looking at the above impulse response, we see the carrier to be

$$h_{\sin}[n] = \frac{\sin[\theta(n+1)]}{\sin(\theta)} u[n] \quad \overset{z}{\longleftrightarrow} \quad H_{\sin}(z) = \frac{1}{1 - 2r\cos(\theta)z^{-1} + r^2z^2}. \tag{5.31}$$

So what would the cos-phase impulse response and it's z-transform look like? Let's try!

$$\begin{aligned} h_{\cos}[n] = \frac{\cos[\theta(n+1)]}{\sin(\theta)} u[n] \qquad \overset{z}{\longleftrightarrow} \qquad H_{\cos}(z) &= \frac{1}{2\sin(\theta)} \left[\frac{e^{j\omega}}{1 - e^{j\omega}z^{-1}} + \frac{e^{-j\omega}}{1 - e^{-j\omega}z^{-1}} \right] \\ &= \frac{1}{2\sin(\theta)} \frac{e^{j\theta}\left(1 - e^{-j\theta}\right) + e^{-j\theta}\left(1 - e^{j\theta}\right)}{(1 - 2\cos(\theta)z^{-1} + z^{-2})} \\ H_{\cos}(z) &= \frac{1}{\sin(\theta)} \frac{\cos(\theta) - z^{-1}}{(1 - 2\cos(\theta)z^{-1} + z^{-2})}. \end{aligned} \tag{5.32}$$

It worked, so we can build the transfer function of our 90-phase shifter to be the fraction of both $H_{\cos}(z)$ and $H_{\sin}(z)$

$$\begin{aligned} H_{90^\circ}(z) &= \frac{H_{\cos}(z)}{H_{\sin}(z)} = \frac{\cos(\theta) - z^{-1}}{\sin(\theta)} \\ &= \cot(\theta) - \frac{z^{-1}}{\sin(\theta)}. \end{aligned} \tag{5.33}$$

We have now found a real good phase shifter $H_{90^\circ}(z)$ that gives really nice results for all kinds of gammatones. The system's zero lies at $z = \frac{1}{\cos(\theta)}$, which describes a maximum phase system.

It is quite interesting to note that this is the *time-reversed* version of the previously simplified transfer function between the sin- and cos-phase 3^{rd} order gammatone, which has a (conjugate) reciprocal zero. It is also possible to derive this relationship from the z-transforms of a causal cosine and an acausal sine. Obviously the acausality of the sine is necessary to obtain $H_{9^\circ 0}(z)$. If two causal sequences are used, the inaccurate (*time-reversed*) version $\tilde{H}_{9^\circ 0}(z)$ will turn out to be the result.

6. Bilinear Transform

The Bilinear Transform can be applied to a continuous-time system's Laplace representation in order to obtain a corresponding discrete-time system representation in the z-domain. Unlike the impulse invariance technique this method doesn't have to cope with aliasing problems.

The Bilinear Transform is defined by

$$s = \frac{2}{T} \frac{1 - z^{-1}}{1 + z^{-1}}, \tag{6.1}$$

it can immediately be plugged into equation 3.8

$$\begin{aligned} G_m(z) &= \frac{(-1)^{m-1}(m-1)!}{2} \cdot \\ &\quad \frac{\left(\frac{2}{T}\frac{1-z^{-1}}{1+z^{-1}}+b-j\omega\right)^m + \left(\frac{2}{T}\frac{1-z^{-1}}{1+z^{-1}}+b+j\omega\right)^m}{\left(\frac{2}{T}\frac{1-z^{-1}}{1+z^{-1}}+b-j\omega\right)^m \left(\frac{2}{T}\frac{1-z^{-1}}{1+z^{-1}}+b+j\omega\right)^m} \\ &= \frac{(-1)^{m-1}(m-1)!}{2}\left(\frac{T}{2}(1+z^{-1})\right)^m \cdot \\ &\quad \frac{\left[1+\frac{T}{2}b-\frac{T}{2}j\omega+\left(\frac{T}{2}b-\frac{T}{2}j\omega-1\right)z^{-1}\right]^m + \left[1+\frac{T}{2}b+\frac{T}{2}j\omega+\left(\frac{T}{2}b+\frac{T}{2}j\omega-1\right)z^{-1}\right]^m}{\left[\frac{1}{\left(1+\frac{2}{T}b\right)^2+\left(\frac{2}{T}\omega\right)^2}\right]^m \left(1-\frac{1+\frac{T}{2}j\omega-\frac{T}{2}b}{1-\frac{T}{2}j\omega+\frac{T}{2}b}z^{-1}\right)^m \left(1-\frac{1-\frac{T}{2}j\omega-\frac{T}{2}b}{1+\frac{T}{2}j\omega+\frac{T}{2}b}z^{-1}\right)^m} \end{aligned} \tag{6.2}$$

The obvious difference between this and the impulse invariance transfer function is that the poles in the transfer function obtained by the bilinear transform ly at $\frac{1\pm\frac{T}{2}j\omega-\frac{T}{2}b}{1\mp\frac{T}{2}j\omega+\frac{T}{2}b}$ whereas the poles in the impulse invariance system are located at $e^{(-b+j\omega)T}$.

7. All-pole and One-zero Gammatone Filters

In order to reduce the computational effort we want to use the All-pole Gammatone transfer functions proposed by Slaney [2] and Lyon [1] for gammatone filtering. By degrading the stop-band attenuation of the gammatone-filters and omitting all zeros in the numerator polynomial, All-pole Gammatone filters (APGF) are obtained. Fortunately APGFs have useful properties, like asymmetry, that give even more appropriate results than the conventional gammatone filters. The desired asymmetry property is met for filter with center frequencies smaller than $\frac{\pi}{2}$, i.e. in gammatone filter bank approaches most of the filters will show up this asymmetry, because of logarithmical frequency spacing.

Describing the APGF's properties an arbitrary second order IIR allpole-filter with the pole-pair $a = re^{j\theta}$ with the z-domain description

$$H_{\mathrm{AP}}(z) = \frac{1}{(1-az^{-1})(1-a^*z^{-1})} \tag{7.1}$$

shall be used. In order to obtain a continuous-time impulse response, we need to use the inverse z-transform. Therefore the partial fraction expansion of the transfer function

$$\begin{aligned} H_{\mathrm{AP}}(z) &= \frac{e^{j\theta}}{e^{j\theta}-e^{-j\theta}}\frac{1}{(1-az^{-1})} - \frac{e^{-j\theta}}{e^{j\theta}-e^{-j\theta}}\frac{1}{(1-a^*z^{-1})} \\ &= \frac{1-r\cos(\theta)z^{-1}}{1-2r\cos(\theta)z^{-1}+r^2z^{-2}} + \cot(\theta)\frac{r\sin(\theta)z^{-1}}{1-2r\cos(\theta)z^{-1}+r^2z^{-2}} \end{aligned} \tag{7.2}$$

can be used to obtain the impulse response

$$\begin{aligned} h_{\mathrm{AP}}[n] &= r^n\frac{\sin[\theta(n+1)]}{\sin(\theta)}u[n] \\ &= r^n\left[\cos(\theta n)+\cot(\theta)\sin(\theta n)\right]u[n]. \end{aligned} \tag{7.3}$$

If we compare equation 7.3 to a first order Gammatone function from equation 4.1 it can be seen that the phase term is now dependent on the center frequency $\phi = \theta - \frac{\pi}{2}$.

It's interesting that this relation can also be expressed in terms of a cos-phase and a sin-phase gammatone.

For an All-pole implementation of a gammatone filter m stages of the above system have to be cascaded, we obtain the APGF

$$G_{\mathrm{AP},m}(z) = \frac{1}{(1 - 2r\cos\theta z^{-1} + r^2 z^{-2})^m}. \tag{7.4}$$

The over-all system impulse response could be derived using the convolution sum on the above impulse response (eq. 7.3), or by partial fraction analysis of the m^{th} order z-domain description (eq. 7.4) and subsequent inverse z-transform, which is in fact not very easy in terms of higher order APGFs.

The One-Zero Gammatone filter (OZGF) has the same form as the APGF, except that there's one zero at $z = 1$ in the numerator polynome. The m^{th} order OZGF is

$$G_{\mathrm{OZ},m}(z) = \frac{1 - z^{-1}}{(1 - 2r\cos\theta z^{-1} + r^2 z^{-2})^m}. \tag{7.5}$$

This additional zero allows a better behaviour of the gammatone filter's lower skirt towards DC, while keeping a computationally compact form. In terms of parallel OZGF banks this zero can be applied in for all channels in front of the filter bank, the parallel filters only need to be APGFs.

8. Polyphase Decomposition, DCT implementation

A polyphase decomposition for the first order Gammatone envelope can be obtained in a simple manner. The first order gammatone envelope is

$$G_{u,1}(z) = \frac{1}{1 - rz^{-1}}. \tag{8.1}$$

For IIR-polyphase decomposition we need to express the denominator polynomials in z^N, where N is the order of the polyphase decomposition.

$$\begin{aligned} G_{u,1}(z) &= \frac{1}{1 - rz^{-1}} = \frac{1 + rz^{-1}}{1 - r^2 z^{-2}} \\ &= \frac{(1 + rz^{-1})(1 + r^2 z^{-2})}{1 - r^4 z^{-4}} = \frac{(1 + rz^{-1})(1 + r^2 z^{-2})(1 + r^4 z^{-4})}{1 - r^8 z^{-8}} \\ &= \frac{(1 + rz^{-1})(1 + r^2 z^{-2})(1 + r^4 z^{-4})(1 + r^8 z^{-8})}{1 - r^{16} z^{-16}} = \dots . \end{aligned} \tag{8.2}$$

For $N = 16$ we can write

$$\begin{aligned} G_{u,1}(z) &= \frac{(1 + rz^{-1})(1 + r^2 z^{-2})(1 + r^4 z^{-4})(1 + r^8 z^{-8})}{1 - r^{16} z^{-16}} = \\ &= \frac{1 + rz^{-1} + r^2 z^{-2} + r^3 z^{-3} + \dots + r^{15} z^{-15}}{1 - r^{16} z^{-16}} \\ &= \frac{1}{1 - r^{16} z^{-16}} + z^{-1}\frac{r}{1 - r^{16} z^{-16}} + \dots + z^{-15}\frac{r^{15}}{1 - r^{16} z^{-16}}, \end{aligned} \tag{8.3}$$

which is a nice IIR polyphase-decomposition, that can be used in front of a 16 point DCT. Using an All-pass transform unequal frequency spacing can be achieved. Unfortunately this is only a first order gammatone, but perhaps a cascade implementation also achieves good results. This form of implementation could be very efficient, but before the resulting filter shapes should be observed.

9. Additional Calculus

In this section we will derive approximations and exact solutions for some of the gammatone's parameters. For example we derived gammatones in the previous sections assuming some bandwidth coefficient B we aren't able to use yet, in fact it wouldn't work as we expect it to do. So one task will be to derive a real relationship between some bandwidth $\Delta\theta_b$ we want the gammatone filter transfer function to have.

In order to implement a gammatone filter bank, we will need to know about the frequency of common magnitude, shared by filters with neighbouring center frequencies, so we will also spend some time deriving an accurate relationship.

9.1. Calculation of the bandwidth.

Assuming that the gammatone's magnitude $|H(e^{j(\theta+\Delta\theta)})|$ around the resonance frequency θ depends only on the local distance from the nearest m-fold pole and some gain factor g, we can use a tangential approximation for the magnitude response

$$|H(e^{j(\theta+\Delta\theta)})| \simeq \frac{g}{\prod_{k=1}^{m}\sqrt{(1-r)^2+\Delta\theta^2}}. \tag{9.1}$$

In order to obtain the -3dB-bandwidth $\Delta\theta_b$ we can constrain the squared magnitude to

$$\begin{aligned}
|H(e^{j\theta})|^2 &\stackrel{!}{=} 2|H(e^{j(\theta+\frac{\Delta\theta_b}{2})})|^2. \qquad (9.2)\\
\frac{g}{\prod_{k=1}^{m}(1-r)^2} &\stackrel{!}{=} \frac{2g}{\prod_{k=1}^{m}\left[(1-r)^2+\left(\frac{\Delta\theta_b}{2}\right)^2\right]}\\
\prod_{k=1}^{m}\sqrt[m]{2}(1-r)^2 &= \prod_{k=1}^{m}\left[(1-r)^2+\left(\frac{\Delta\theta_b}{2}\right)^2\right]\\
\Longrightarrow \qquad r &= 1-\frac{\Delta\theta_b}{2\sqrt{\sqrt[m]{2}-1}}\\
\text{or the other way round} \qquad \Delta\theta_b &= 2(1-r)\sqrt{\sqrt[m]{2}-1} \qquad (9.3)
\end{aligned}$$

With this simple relation it's easy to control the gammatone filter's bandwidth, the results are quite accurate for small bandwidths.

9.2. Calculation of the overlap frequency.

9.2.1. *Simple solutions.* Neighbouring filters' frequency of common magnitude shall be derived here. One could simply move the lower filter's half bandwidth up from its center frequency, and the higher filter's half bandwidth down from its center frequency. Building the mean value of these two gives our first, inaccurate frequency

$$\omega = \frac{\theta_i + \frac{1}{2}\Delta\theta_{i,b} + \theta_{i+1} - \frac{1}{2}\Delta\theta_{i+1,b}}{2}. \tag{9.4}$$

Unfortunately this doesn't really match, so one could also try logarithmic relations like

$$\omega = \sqrt{\theta_i\sqrt{\Delta\theta_{i,b}}\cdot\frac{\theta_{i+1}}{\sqrt{\Delta}\theta_{i+1,b}}}. \tag{9.5}$$

This works better, but as well doesn't fit very accurate. So we need to find more accurate methods.

9.2.2. *Solution for the analytic APGF.* We approximate the i^{th} gammatone's magnitude $|H_i(e^{j\omega})|$ by assuming it depends only on the distance to the nearest (m-fold) pole $p_i = r_i e^{j\theta_i}$. Now we want to find the frequency at which two neighbouring filters $|H_1(e^{j\omega})|$ and $|H_2(e^{j\omega})|$ have the same magnitude, i.e. the overlapping point of the two magnitude functions. *(In case of analytic All-pole gammatone filters, this would be the exact solution!)*
For this aim we can constrain the two magnitudes to be

$$\begin{aligned} |H_1(e^{j\omega})|^2 &\stackrel{!}{=} |H_1(e^{j\omega})|^2 \\ \Longrightarrow \frac{(1-r_1)^{2m}}{\left|e^{j\omega} - r_1 e^{j\theta_1}\right|^{2m}} &= \frac{(1-r_2)^{2m}}{\left|e^{j\omega} - r_2 e^{j\theta_2}\right|^{2m}} \end{aligned} \tag{9.6}$$

$$\begin{aligned} (1-r_2)^2 \left\{ [\cos(\omega) - \Re\{p_1\}]^2 + [\sin(\omega) - \Im\{p_1\}]^2 \right\} = \\ \left\{ [\cos(\omega) - \Re\{p_2\}]^2 + [\sin(\omega) - \Im\{p_2\}]^2 \right\} (1-r_1)^2 \end{aligned}$$

$$\begin{aligned} (1-r_2)^2 - 2\Re\{p_1\}(1-r_2)^2\cos(\omega) - 2\Im\{p_1\}(1-r_2)^2\sin(\omega) + r_1^2(1-r_2)^2 = \\ (1-r_1)^2 - 2\Re\{p_2\}(1-r_1)^2\cos(\omega) - 2\Im\{p_2\}(1-r_1)^2\sin(\omega) + r_2^2(1-r_1)^2 \end{aligned}$$

$$\begin{aligned} \left[\Re\{p_2\}(1-r_1)^2 - \Re\{p_1\}(1-r_2)^2\right]\cos(\omega) + \\ \left[\Im\{p_2\}(1-r_1)^2 - \Im\{p_1\}(1-r_2)^2\right]\sin(\omega) = \ldots \\ \frac{(1-r_1)^2(1+r_2^2) - (1-r_2)^2(1+r_1^2)}{2} \end{aligned}$$

$$\mathrm{C}\cos(\omega) + \mathrm{S}\sin(\omega) = \mathrm{B} \qquad \Longrightarrow \qquad \mathrm{A}\cdot\sin(\omega + \phi) = \mathrm{B} \tag{9.7}$$

$$\begin{aligned} \mathrm{C} &= \Re\{p_2\}(1-r_1)^2 - \Re\{p_1\}(1-r_2)^2 \\ \mathrm{S} &= \Im\{p_2\}(1-r_1)^2 - \Im\{p_1\}(1-r_2)^2 \\ \mathrm{A} &= \sqrt{S^2 + C^2} \\ \phi &= \angle(C, S) \\ \mathrm{B} &= \frac{(1-r_1)^2(1+r_2^2) - (1-r_2)^2(1+r_1^2)}{2} \\ \omega &= 3\pi(1 \pm_1 1) \pm_1 \arcsin\left(\frac{B}{A}\right) - \phi \end{aligned} \tag{9.8}$$

For real-valued gammatone filters the results are not good enough, because of the complex conjugate pole, the influence of which can't be neglected in the magnitude function.

9.2.3. *Solution for the real-valued APGF.* Now we want to do the same as above for a real-valued transfer function with a pair of complex conjugate poles $p_i = r_i e^{\pm j\theta_i}$.

First of all we want to simplify the squared magnitude function $|H_i(e^{j(\omega)})|^2$

$$
\begin{aligned}
|H_i(e^{j\omega})|^2 &= \frac{g_i^{2m}}{\left|e^{j\omega} - r_1 e^{j\theta_i}\right|^{2m} \left|e^{j\omega} - r_i e^{-j\theta_i}\right|^{2m}} \\
&= \frac{g_i^{2m}}{\left[(\cos(\omega) - \Re\{p_i\})^2 + (\sin(\omega) - \Im\{p_i\})^2\right]^m \left[(\cos(\omega) - \Re\{p_i\})^2 + (\sin(\omega) + \Im\{p_i\})^2\right]^m} \\
&= \frac{g_i^{2m}}{\left\{\left[(1 + r_i^2 - 2\Re\{p_i\}\cos(\omega)) - 2\Im\{p_i\}\sin(\omega)\right]\left[(1 + r_i^2 - 2\Re\{p_i\}\cos(\omega)) + 2\Im\{p_i\}\sin(\omega)\right]\right\}^m} \\
&= \frac{g_i^{2m}}{\left((1 + r_i^2)^2 - 4\Re\{p_i\}(1 + r_i^2)\cos(\omega) + 4r_i^2\cos^2(\omega) - 4\Im^2\{p_i\}\right)^m} \\
&= \frac{g_i^{2m}}{4^m\left(r_i^2\cos^2(\omega) - (1 + r_i^2)\Re\{p_i\}\cos(\omega) + \Re\{p_i\} + \frac{(1 - r_i^2)^2}{4}\right)^m}
\end{aligned}
$$

Now we can solve our magnitude constraint

$$|H_1(e^{j\omega})|^2 \quad \stackrel{!}{=} \quad |H_1(e^{j\omega})|^2$$

$$
\begin{aligned}
g_2^2\left(r_1^2\cos^2(\omega) - (1 + r_1^2)\Re\{p_1\}\cos(\omega) + \Re\{p_1\} + \frac{(1 - r_1^2)^2}{4}\right) = \\
g_1^2\left(r_2^2\cos^2(\omega) - (1 + r_2^2)\Re\{p_2\}\cos(\omega) + \Re\{p_2\} + \frac{(1 - r_2^2)^2}{4}\right)
\end{aligned}
$$

$$
\begin{aligned}
&\left[g_2^2 r_1^2 - g_1^2 r_2^2\right]\cos^2(\omega) + \\
&\qquad \left[g_1^2(1 + r_2^2)\Re\{p_2\} - g_2^2(1 + r_1^2)\Re\{p_1\}\right]\cos(\omega) + \\
&\qquad g_2^2\left(\Re\{p_1\} + \frac{(1 - r_1^2)^2}{4}\right) - g_1^2\left(\Re\{p_2\} + \frac{(1 - r_2^2)^2}{4}\right) = 0 \qquad (9.9)
\end{aligned}
$$

$$
\begin{aligned}
0 &= \mathrm{A}\cos(\omega) + \mathrm{B}\cos(\omega) + \mathrm{C} \\
\omega &= \arccos\left[\frac{1}{2\mathrm{A}}\left(-\mathrm{B} \pm \sqrt{\mathrm{B}^2 - 4\mathrm{AC}}\right)\right] \qquad (9.10)
\end{aligned}
$$

For two APGF transfer functions this is the exact solution. The real roots describe the existing frequencies of equal magnitude, speaking of positive normalized frequencies within $[0 \ldots 2\pi]$.

9.3. **Gain normalization.** In this section simple gain normalization formulae are calculated avoiding the effort of polynomial root solving.

9.3.1. *The APGF gain factor.* The gain factors for normalized APGFs $g_{i,\mathrm{APGF}}$ can be computed for the APGF in its center frequency

$$
\begin{aligned}
g_{i,\mathrm{APGF}} &= \frac{1}{|H(e^{j\theta_i})|} \\
&= \left|\left(1 - r_i e^{j\theta_i} e^{-j\theta_i}\right)\left(1 - r_i e^{-j\theta_i} e^{-j\theta_i}\right)\right| \\
&= (1 - r_i)\sqrt{\left[1 - r_i\cos(2\theta_i)\right]^2 + r_i^2\sin^2(2\theta_i)} \\
&= (1 - r_i)\sqrt{1 - 2r_i\cos(2\theta_i) + r_i^2}. \qquad (9.11)
\end{aligned}
$$

9.3.2. *The OZGF gain factor.* In order to get more appropriate results for the OZGF we propose a modified gain factor $g_{i,\mathrm{OZGF}}$ that takes the zero at $z = 1$ into account

$$\begin{aligned} g_{i,\mathrm{OZGF}} &= \frac{g_{i,\mathrm{APGF}}}{\left(\sqrt{(1-\cos(\theta_i)^2) + \sin^2(\theta_i)}\right)^{\frac{1}{m}}} \\ &= \frac{g_i}{(2-2\cos(\theta_i))^{\frac{1}{2m}}}. \end{aligned} \tag{9.12}$$

In terms of the OZGF this approximation does really a good job for finding the overlapping point in the magnitude responses of two neighbouring filters.

9.3.3. *The GF gain factor.* For the GF we propose to take in account that there are m zeros on the real axis. For high frequencies those zeros move towards $z = 1$ and for low frequencies towards $z = -1$. We empirically found some modification for the gain factor $g_{i,\mathrm{GF}}$ that fits approximately for all gammatone orders we want to use

$$g_{i,\mathrm{GF}} = \frac{g_{i,\mathrm{APGF}}}{(2-2\cos(\theta_i))^{\frac{1}{5}}(2+2\cos(\theta_i))^{\frac{1}{4}}}. \tag{9.13}$$

9.4. **Calculation of the overlap phase.** The filter overlap phase between two neighbouring filters gives information about probable cancellations when summing up the two bandpass signals. In order to avoid such cancellations one could try to exactly match the phase between the two filters at the overlap frequency, though this is hard to acheive, or one could look if a simple sign change can already help to avoid major cancellations. We make an approximation by using the Laplace-domain description of a single-pole filter to describe a complex conjugate pole pair. From the transfer function

$$H(s) = \frac{b}{s + b - j\omega_c} \tag{9.14}$$

we can write for the band edge magnitude a of the m-fold cascaded filter at $s = j\omega$

$$\begin{aligned} |H(s)|^m &\stackrel{!}{=} a \\ \frac{b^m}{\prod_{k=1}^{m}\sqrt{(\omega-\omega_c)^2+b^2}} &= a \\ \frac{b^m}{\prod_{k=1}^{m}\sqrt{\Delta\omega_{casc}^2+b^2}} &= a \\ \frac{b}{\sqrt{\Delta\omega_{casc}^2+b^2}} &= \sqrt[m]{a} \\ \Delta\omega_{casc} &= b\sqrt{\frac{1}{\sqrt[m]{a^2}} - 1}. \end{aligned} \tag{9.15}$$

A convenient expression can be achieved by bringing the bandwidth of the single stage filter in relation to the cascaded filter's bandwidth, thus for the single stage

filter the bandwidth is

$$\begin{aligned} |H(s)| &\stackrel{!}{=} a \\ \frac{b}{\sqrt{\Delta\omega^2 + b^2}} &= a \\ \Delta\omega &= b\sqrt{\frac{1}{a^2} - 1}. \end{aligned} \tag{9.16}$$

Finally we can express the phase of the cascaded filter stages at $s = j(\omega_c + \Delta\omega_{casc})$

$$\begin{aligned} \angle\{H^m(s)\} &= m \cdot \arctan\left[H(s)\right] \\ &= m \cdot \arctan\left[\frac{\Delta\omega}{b}\right]_{\omega=\omega_c+\Delta\omega_{casc}} \\ &= m \cdot \arctan\left[\frac{\omega_c + \Delta\omega_{casc} - \omega_c}{b}\right] \\ &= m \cdot \arctan\left[\frac{b\sqrt{\frac{1}{\sqrt[m]{a^2}} - 1}}{b}\right] \\ \Longrightarrow \angle\{H^m(s)\} &= m \cdot \arctan\left[\sqrt{\frac{1}{\sqrt[m]{a^2}} - 1}\right] \end{aligned} \tag{9.17}$$

From this we can find that the angle between two neighbouring filters at their common band edge

$$\begin{aligned} \Delta\angle\{H_{k,k+1}(s)\} &= \angle\{H_k^m(s)\}_{\omega=\omega_c+\Delta\omega_{casc}} - \angle\left\{H_{k+1}^m(s)\right\}_{\omega=\omega_c-\Delta\omega_{casc}} \\ \Longrightarrow \Delta\angle\{H_{k,k+1}(s)\} &= 2m \cdot \arctan\left[\sqrt{\frac{1}{\sqrt[m]{a^2}} - 1}\right]. \end{aligned} \tag{9.18}$$

A simple rule for sign alternation can be found to be

$$f_{sign}[H_{k,k+1}(s)] = \begin{cases} 1 & \text{, if } \left[2m \cdot \arctan\left(\sqrt{\frac{1}{\sqrt[m]{a^2}} - 1}\right)\right]_{\text{mod } \pi} < \frac{\pi}{2} \\ -1 & \text{, if } \left[2m \cdot \arctan\left(\sqrt{\frac{1}{\sqrt[m]{a^2}} - 1}\right)\right]_{\text{mod } \pi} > \frac{\pi}{2} \end{cases} \tag{9.19}$$

For $a = 10^{-\frac{|\text{overlapdB}|}{10}}$ one could also write

$$f_{sign}\left[H_{k,k+1}(s)\right] = \text{sign}\left\{\cos\left[\Delta\angle\left(H_{k,k+1}(s)\right)\right]\right\} \tag{9.20}$$

References

[1] Richard F. Lyon. The all-pole gammatone filter and auditory models. Apple Computer, 1993.

[2] Malcolm Slaney. An efficient implementation of the patterson-holdsworth auditory filter bank. Technical report #35, Apple Computer, 1993. Perception Group-Advanced Technology Group.