Institut für Nachrichtengeräte
und Datenverarbeitung
Prof. Dr.-Ing. Peter Vary

RWTH AACHEN UNIVERSITY

Modellbasierte psychoakustische Störgeräuschreduktion

Autor: Riko Brandes (riko.brandes@gmail.com)

Inhaltsverzeichnis

Kapitel 1

Einleitung

Heutzutage wird zunehmend über Sprachkommunikationssysteme kommuniziert. Durch die mobile Anwendung sind Gespräche an allen erdenklichen Orten der Welt möglich. Immer häufiger werden Telekonferenzsysteme genutzt, um Konferenzen über Kontinente hinweg abzuhalten. Dadurch ergeben sich Szenarien, welche durch akustisch gestörte Umgebungen die Sprachverständlichkeit herabsetzten können. Gesprächspartner A kann aufgrund des beim Gesprächspartner B eingespeisten Störgeräuschs diesen nicht mehr gut verstehen.

Zur Erhöhung der Sprachverständlichkeit werden Störgeräuschreduktionssysteme eingesetzt. Das Ziel aller Verfahren zur Störgeräuschreduktion ist es, gleichzeitig das gestörte Sprachsignal möglichst gut von der vorhandenen Störung zu befreien, während die damit verbundene Sprachverzerrung des eigentlichen Nutzsignals so gering wie möglich gehalten wird. Der kritische Teil einer Störgeräuschreduktion besteht in der Schätzung der zeitveränderlichen spektralen Zusammensetzung der Störung. Auf Basis der Schätzung des Störspektrums werden Gewichtsfaktoren berechnet, mit denen das gestörte Eingangssignal frequenz- und zeitadaptiv gefiltert wird.

Fehler bei der Schätzung des Störspektrums führen direkt zu deutlich wahrnehmbaren Artefakten im gefilterten Signal. Als besonders störend wird das sogenannte *Musical Noise* empfunden. Dies wird durch hohe Fehler bei der Schätzung, welche vorwiegend bei geringem Signal- zu Störabstand auftreten, hervorgerufen. Letzterer ist besonders für höhere Frequenzbereiche aufgrund der Energieverteilung von Sprache sehr klein. Außerdem wird durch die spektrale Gewichtung mittels der konventionellen Störgeräuschreduktion die Sprache so stark gedämpft, dass das Klangbild als verzerrt wahrgenommen wird, oder im Extremfall die Sprachverständlichkeit signifikant verringert.

Das Auftreten von Artefakten und die Verzerrung der Sprache durch unnötig hohe Sprachdämpfung soll in dieser Arbeit unter Verwendung psychoakustischer Eigenschaften des menschlichen Gehörs in einer zweistufigen Störgeräuschreduktion möglichst verringert werden. Dazu werden zunächst verschiedene psychoakustische Modelle zur Verwendung im Störgeräuschreduktionssystem angepasst und erweitert. Das weitere Vorgehen besteht in der Entwicklung neuer psychoakustischer Filterregeln, welche in Kombination mit dem psychoakustischen Modell das resultierende Klangbild des gefilterten Sprachsignals weiter verbessern und die Sprachverständlichkeit erhöhen soll.

1.1 Aufbau der Arbeit

In Kapitel 2 wird kurz thematisch in die konventionelle Störgeräuschreduktion eingeführt. Eine Beschreibung der spektralen Energieverteilung von Sprache leitet zur Psychoakustik über. Nachfolgend werden die zur Bewertung der Leistung von Störgeräuschreduktionssystemen wichtigen instrumentellen Maße vorgestellt. Die Probleme der konventionellen Störgeräuschreduktion werden beleuchtet und daraus die Motivation für eine psychoakustisch basierte zweistufige Störgeräuschreduktion abgeleitet.

In Kapitel 3 werden die wesentlichen Aspekte der Psychoakustik zur Nutzung für die Störgeräuschreduktion behandelt. Dazu werden Eigenschaften des menschlichen Gehörs als System mathematisch beschrieben. Die für die psychoakustische Störgeräuschreduktion besonders interessanten Phänomene wie die spektrale und temporale Verdeckung und die Empfindung von Lautstärke werden ausführlich erläutert. Das Kapitel schließt mit einer Anforderungsliste an die in der zweistufigen Störgeräuschreduktion zu verwendenden psychoakustischen Modelle zusammenfassend ab.

Letztere werden in Kapitel 4 behandelt. Hier wird auf nötige Anpassungen und sinnvolle Erweiterungen der psychoakustischen Modelle eingegangen und abschließend eine Bewertung dieser in Bezug auf die in Kapitel 3 gestellten Anforderungen vorgenommen.

Das fünfte Kapitel beschäftigt sich mit den psychoakustischen Filterregeln als Teil der zweiten Stufe der zweistufigen Störgeräuschreduktion. Angeknüpft an eine vorhandene Gewichtungsregel, wird diese optimiert und neue Filterregeln entwickelt, welche die in Kapitel 4 vorgestellten psychoakustischen Modelle verwenden. Am Ende des Kapitels werden die Filtervarianten kategorisiert.

Kapitel 6 widmet sich den Ergebnissen der psychoakustisch basierten Störgeräuschreduktion. Mittels mehrerer Benchmarks werden die besten Kombinationen aus Filterregeln und psychoakustischen Modellen im zweistufigen Störgeräuschreduktionssystem untersucht. Letzteres wird mit der konventionellen Störgeräuschreduktion unter Heranziehung ausgewählter instrumenteller Maße aus Kapitel 2 und spektraler Betrachtungen verglichen. Dabei werden die Aspekte der Sprach- und Störgeräuschdämpfung, Sprachverzerrungen und des Auftretens von „Musical Noise“ behandelt. Das Kapitel schließt mit einem Sprachverständlichkeitsvergleich und einer Zusammenfassung der Eigenschaften verschiedener Ansätze ab. Zusammenfassung und Ausblick dieser Arbeit finden sich Kapitel 7.

Kapitel 2

Grundlagen

2.1 Konventionelle Störgeräuschreduktion

Das verrauschte Signal setzt sich aus dem Sprachsignal $s(t)$ und dem Rauschsignal $n(t)$ zusammen:

$$x(t) = s(t) + n(t) \tag{2.1}$$

bzw. im Frequenzbereich nach Anwendung einer Fouriertransformation:

$$X(\mu, \lambda) = S(\mu, \lambda) + N(\mu, \lambda) \tag{2.2}$$

$X(\mu, \lambda)$ repräsentiert die Fouriertransformierte des verrauschten Signals $x(t)$. Die Indizes μ und λ kennzeichnen die Frequenzlinie und den Rahmen. Geschätzte Größen werden mit einem „Dach“ über dem Variablenbuchstaben versehen. Kleine Buchstaben repräsentieren in der Regel Größen im Zeitbereich, große Buchstaben Größen im Frequenzbereich.

Voraussetzung zur Verwendung von Filterregeln wie z. B. des Wiener Filters ist es, dass das zu verarbeitende Signal stationär bzw. quasistationär ist. Dies wird durch Segmentierung des Signals in Segmente (oder Rahmen) gewährleistet. Ein Rahmen enthält Abschnitte des Zeitsignals von ungefähr 20 ms Länge, in dem Sprache als quasistationär betrachtet werden kann.

2.1.1 Systembeschreibung

Das allgemeine Blockschaltbild einer konventionellen Störgeräuschreduktion ist in Abbildung 2.1 dargestellt. Es wird nur die spektrale Gewichtung betrachtet.

Zur Berechnung der Filterregel, welche auf das verrauschte Signal angewendet wird, muss der Störabstand geschätzt werden. Dafür sind Schätzungen der Amplituden des Störsignals $\hat{N}(\mu, \lambda)$ und des verrauschten Signals $X(\mu, \lambda)$ nötig. Ersteres kann z. B. mit einem sogenannten „SPP Tracker“ [8] gewonnen werden. Die resultierende Filterregel $H(\mu, \lambda)$ wird mit dem Spektrum des verrauschten Signals multipliziert, was einer Faltung des verrauschten Signals $x(t)$ mit der Impulsantwort $h(t)$ entspricht.

$$\hat{S}(\mu, \lambda) = H(\mu, \lambda) \cdot X(\mu, \lambda) \tag{2.3}$$

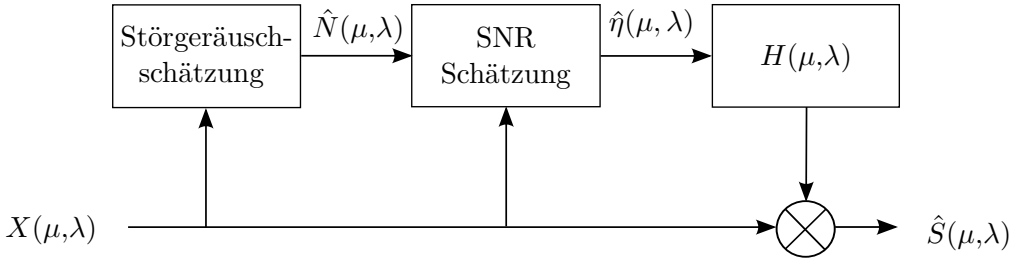


Abbildung 2.1: Blockschaltbild der konventionellen Störgeräuschreduktion.

Das im Blockschaltbild verwendete Spektrum X des Rauschsignals wird mittels Fensterung der Segmente des zeitdiskreten Signals $x(k)$ und anschließender Transformation in den Frequenzbereich gewonnen. Die Rücktransformation des verbesserten Spektrums $\hat{S}$ erfolgt analog unter zusätzlicher Verwendung des sogenannten „overlap add“ Verfahrens. Eine gute Beschreibung dazu findet sich in [23].

2.2 Spektrale Energieverteilung von Sprache

Untersuchungen einer Testmenge von 84 Sprachsignalen aus der NNT Sprachdatenbank mit einer Abtastrate von 16 kHz mit gleichen Anteilen von weibl. und männl. Sprechern [16] haben gezeigt, dass 99 % der Energie im Sprachsignal im Frequenzbereich von 0 bis 3,4 kHz enthalten ist und nur 1 % im darüber liegenden Frequenzbereich bis 8 kHz.

Die Berechnung der Energieverteilung über den Frequenzbereich von 92 Störsignalen der NOISEX Datenbank ergibt, dass die Energie des Rauschens sich auf nur 89 % im unteren betrachteten Frequenzbereich (0 bis 3,4 kHz) und 11 % im oberen Frequenzbereich 3,4 kHz bis 8 kHz aufteilt. Für die sich damit ergebenden Störabstände in dem oberen Frequenzbereich führt die darauf basierende Rauschunterdrückung zu einer starken Dämpfung dieser Sprachanteile, welche für die Natürlichkeit und Verständlichkeit von großer Bedeutung sind.

2.3 Schalldruckpegel

Der Schalldruckpegel *(engl. sound pressure level (SPL))* wird in *Dezibel* angegeben. Der Schalldruck ist dabei nach internationaler Konvention für das Medium Luft auf 20 μPascal referenziert.

$$SPL = 20 \cdot log_{10}(\frac{p_{rms}}{20\mu\text{Pascal}}) \quad [21] \tag{2.4}$$

p_{rms} repräsentiert den Effektivwert des Drucks in μPascal.

2.4 Instrumentelle Maße

Zur objektiven Bewertung der Sprachqualität des mittels Störgeräuschreduktion gewonnenen Signals werden folgende Maße verwendet [13].

2.4.1 Cepstrale Distanz

Das Cepstrum eines Signals $x(k)$ im Zeitbereich wird aus der diskreten inversen Fouriertransformation (IDFT) [1] des logarithmierten Betragsspektrums gebildet. Die Berechnung wird für jeden Rahmen λ und für die Frequenzlinien μ einzeln durchgeführt.

$$C_x^{(\lambda)}(\mu) = IDFT\left\{ln|DFT\left\{x(\lambda,k)\right\}|\right\} \ , \quad \mu \in [0,1,...,M-1] \tag{2.5}$$

Mit M Cepstralkoeffizienten kann das mittels diskreter Fouriertransformation gewonnene Betragsspektrum $X(\mu)$ des Signals $x(k)$ vollständig dargestellt werden, während für die grobe spektrale Struktur $N_{CD} = 32 < M$ Koeffizienten bei einer Abtastfrequenz von $fs = 16\,\text{kHz}$ ausreichen [16]. Die hier betrachtete cepstrale Distanz ist die Distanz zwischen den Cepstren des originalen Sprachsignals s (nicht dessen Schätzwert $\hat{s}$) und des gefilterten originalen Sprachsignals $\tilde{s}$ [13].

$$CD(\lambda) = \frac{10}{ln(10)} \sqrt{[C_s^{(\lambda)}(0) - C_{\tilde{s}}^{(\lambda)}(0)]^2 + 2 \cdot \sum_{i=1}^{N_{CD}} [C_s^{(\lambda)}(i) - C_{\tilde{s}}^{(\lambda)}(i)]^2} \tag{2.6}$$

Die gemittelte cepstrale Distanz basiert auf der Berechnung aller Mittelwerte $CD(\lambda)$ der Rahmen, welche Sprachaktivität aufweisen.

$$CD = \frac{1}{C(\mathbf{M_s})} \sum_{\lambda \in \mathbf{M_s}} CD(\lambda) \ , \ \text{in [dB]} \tag{2.7}$$

Der Vektor $\mathbf{M_s}$ enthält alle zu berücksichtigenden Rahmen [13]. Die cepstrale Distanz gewichtet Unterschiede in den Sprachsignalen bei niedrigen Frequenzen aufgrund der Verwendung des natürlichen Logarithmus stärker als ein lineares Maß. Bei Berücksichtigung der oben erwähnten spektralen Energieverteilung von Sprache wird klar, dass dieses Maß besonders den Eigenschaften von Sprache gerecht wird.

2.4.2 Segmentelle Sprachdämpfung

Das Verhältnis der Energien des Originalsignals s und dessen gefilterten Signals $\tilde{s}$ wird als Sprachdämpfung bezeichnet. Berechnet wird diese für jeden Rahmen einzeln (segmentell):

$$SegSA = 10 \log_{10} \left(\frac{1}{C(\mathbf{K_s})} \sum_{k \in \mathbf{K_s}} \frac{E\left\{s^2(k)\right\}}{E\left\{\tilde{s}^2(k)\right\}} \right) \ , \ \text{in [dB]} \tag{2.8}$$

$\mathbf{K_s}$ ist die Menge aller Abtastwerte. $C(\mathbf{K_s})$ repräsentiert die Anzahl der Abtastwerte. Verglichen mit der cepstralen Distanz sind bei diesem Maß die Verzerrungen für alle Frequenzbereiche gleich „gewichtet“. Daher gibt das Maß keinen Aufschluss darüber, in welcher Art und Weise die Sprache verzerrt wird. Das Sprachdämpfungsmaß $SegSA$ ist besonders in Kombination mit der Störgeräuschdämpfung $SegNA$ (siehe unten) interessant, da die Differenz beider Größen $NA - SA$ die effektive Störgeräuschdämpfung beschreibt [13].

[1] IDFT: Inverse Diskrete Fouriertransformation, DFT: diskrete Fouriertransformation

2.4.3 Segmentelle Störgeräuschdämpfung

Analog zur segmentellen Sprachdämpfung ist die Störgeräuschdämpfung NA definiert. Es stellt die Energie des ursprünglichen Rauschsignals n mit dessen gefilterten Variante $\tilde{n}$ ins Verhältnis.

$$NA = 10\log_{10}\left(\frac{1}{C(\mathbf{K_n})}\sum_{k\in\mathbf{K_n}}\frac{E\left\{n^2(k)\right\}}{E\left\{\tilde{n}^2(k)\right\}}\right) \quad , \text{ in [dB]} \tag{2.9}$$

2.4.4 Differenz der segmentellen Störgeräusch- und Sprachdämpfung

Die Differenz zwischen der segmentellen Störgeräuschdämpfung $SegNA$ und der segmentellen Sprachdämfpung SA ist in Bezug auf die Eigenschaften der spektralen Gewichtung interessant, da meistens mit einer hohen Störgeräuschdämpfung auch eine ungewünscht hohe Sprachdämpfung einhergeht. Ein Filteralgorithmus lässt sich also an der Höhe der Differenz der beiden Maße anschaulich messen. Mit steigender Differenz ist die Filterung bzgl. Sprach- und Störanteil selektiver, was sich in einem besseren Klangbild des resultierenden gefilterten Signals zeigt.

$$SegDA = SegNA - SegSA \tag{2.10}$$

2.4.5 Segmenteller Störabstand

Der segementelle Sprachstörabstand $SegSNR$ basiert auf dem Mittelwert des Verhältnisses der Energie des originalen Sprachsignals s und der Energie der als Störsignalsignal aufgefassten Differenz zwischen originalem und verbessertem Sprachsignal $(s-\hat{s})$ [13]. Für den Störabstand eines einzelnen Rahmens λ mit N Abtastwerten gilt:

$$SegSNR(\lambda) = 10\log_{10}\left(\frac{\sum_{\nu=0}^{N-1}s^2(\nu+\lambda N)}{\sum_{\nu=0}^{N-1}(s(\nu+\lambda N)-\hat{s}(\nu+\lambda N))^2}\right) \quad , \text{ in [dB]} \tag{2.11}$$

Die untere Grenze des Wertebereichs des Störabstands wird auf $-20\,\text{dB}$ begrenzt. Der Mittelwert des segmentellen Störabstands ergibt sich aus der Mittelung über die Menge $\mathbf{M_s}$ der Rahmen, welche Sprachaktivität enthalten.

$$SegSNR = \frac{1}{C(\mathbf{M_s})}\sum_{\lambda\in\mathbf{M_s}}SegSNR(\lambda) \quad , \text{ in dB} \tag{2.12}$$

2.4.6 Segmenteller Sprachstörabstand

Analog zum segmentellen Störabstand wird der Sprachstörabstand berechnet. Anstatt des verbesserten Signals $\hat{s}$ wird das gefilterte Signal $\tilde{s}$ in die Differenz eingesetzt.

$$SegSpSNR(\lambda) = 10\log_{10}\left(\frac{\sum_{\nu=0}^{N-1}s^2(\nu+\lambda N)}{\sum_{\nu=0}^{N-1}(s(\nu+\lambda N)-\tilde{s}(\nu+\lambda N))^2}\right) \tag{2.13}$$

Der globale Wert ergibt sich aus dem Mittelwert über alle Rahmen.

$$SegSpSNR = \frac{1}{C(\mathbf{M_s})} \sum_{\lambda \in \mathbf{M_s}} SegSpSNR(\lambda) \tag{2.14}$$

Je niedriger der segmentelle Sprachstörabstand ist, desto stärker wird die Sprache durch die spektrale Gewichtung verzerrt. Laut [16] weist dieses Maß gegenüber dem globalen Störabstand eine deutlich höhere Korrelation mit Ergebnissen auditiver Studien auf.

2.4.7 Objektive psychoakustische Bewertungsmaße

Die Bewertung von Sprache durch Algorithmen, welche das menschliche Gehör [2] einschließlich der Verarbeitung im Gehirn nachempfinden, bilden das Bindeglied zwischen den oben vorgestellten instrumentellen Maßen und zeitaufwändigen Hörtests. Solche Algorithmen werden als psychoakustische Bewertungsmaße bezeichnet, da sie die Psychoakustik, also die Wahrnehmung der Akustik messen. Die Bewertung der Verständlichkeit bzw. der Sprachqualität findet auf einer Skala von eins (schlecht) verständlich bis fünf (gut verständlich) statt, bzw. von Null (schlecht) bis Eins (sehr gut). Zwei Algorithmen werden in dieser Arbeit verwendet. Der PESQ Standard [3] evaluiert Sprachsignale bis Bandbreiten von 4 kHz, die Breitbandversion (wideband PESQ) bis 8 kHz. Der zweite Standard ist ein objektives Kurzzeit Verständlichkeitsmaß (STOI)[4]. Laut Untersuchungen von Taal [35] ist es zur Messung der Sprachverständlichkeit vor und nach der Filterung von allen betrachteten Maßen [5] am nächsten an Hörtestergebnissen. Für die Anwendung beider Maße muss das Signal unterabgetastet werden, da die Abtastraten im Bereich von 32 bis 48 kHz liegen.

2.5 Referenzstörgeräuschreduktion

Die in dieser Arbeit verwendete Referenzstörgeräuschreduktion mit der alle psychoakustisch basierten Störgeräuschreduktionen verglichen werden, nutzt einen Wiener Filter als Gewichtungsregel. Die Funktionsweise wird im folgenden beschrieben.

2.5.1 Gewichtungsregel Wiener Filter

Das Wiener Filter verwendet den Störabstand zur Berechnung der Filterregel. Die Regel minimiert den quadratischen Fehler zwischen dem Schätzsignal $S(\hat{\mu}, \lambda)$ und einem Referenzsignal. Für die Herleitung sei auf [31] verwiesen. Der Störabstand wird mittels des sogenannten „Decision-Directed“ Ansatz berechnet. Dieser wiederum bedient sich einer Störgeräuschschätzung nach [8] einem sogenannten „SPP Tracker“ [6].

[2] unter dem Begriff ist sowohl das Ohr als auch die neuronale Ebene (Gehirn) gemeint

[3] Perceptual Evaluation of Speech Quality (Bewertung der Sprachqualität in Bezug auf die Wahrnehmung

[4] STOI: short-time objective intelligibility measure

[5] betrachtet wurden: SegSNR, log-likelihood ratio, Itakura-Saito, CD, Weighted-Spectral Slope Metric (WSS), Normalized Frequency Weighted SSNR (FWS), PESQ, Dau auditory model (DAU), coherence SII (CSII), covariance based STI (CSTI), STOI

[6] SPP: speech presence probability

Das Filtergewicht für die betrachtete Frequenzlinie μ des Rahmens λ berechnet sich aus den Erwartungswerten der spektralen Leistungsichten des klaren Sprachsignals S und des Störanteils N.

$$H(\mu,\lambda) = \frac{E\{|S(\mu,\lambda)|^2\}}{E\{|S(\mu,\lambda)|^2\} + E\{|N(\mu,\lambda)|^2\}} \tag{2.15}$$

Da die Größen $E\{|S(\mu,\lambda)|^2\}$, $E\{|N(\mu,\lambda)|^2\}$ in der Praxis nicht verfügbar sind, werden sie geschätzt.

$$H_(\mu,\lambda) = \frac{|\hat{S}(\mu,\lambda)|^2}{|\hat{S}(\mu,\lambda)|^2 + |\hat{N}(\mu,\lambda)|^2} \tag{2.16}$$

Durch Umformen erhält man eine vom geschätzten Störabstand $\hat{\eta}(\mu,\lambda)$ abhängige Größe. Die Berechnung wird in den folgenden Unterabschnitten erklärt.

$$H(\mu,\lambda) = \frac{\hat{\eta}(\mu,\lambda)}{1+\hat{\eta}(\mu,\lambda)} \tag{2.17}$$

Das menschliche Ohr ist relativ unempfindlich gegenüber Störungen in der Phase im Sprachsignal und nur bei niedrigen Signal zu Rauschverhältnissen fällt die Amplitude des Rauschanteils $N(\mu,\lambda)$ relativ zur Amplitude der Sprache $S(\mu,\lambda)$ so groß aus, dass eine Phasenänderung beim Vergleich von $arg(S(\mu,\lambda))$ und $arg(X(\mu,\lambda))$ wahrgenommen wird [13].

Für das geschätzte Sprachsignal wird die Phase des verrauschten Signals angenommen. Die Gewichte der Gewichtungfunktion H sind daher reell und größer Null, damit die Phase der Schätzung des Sprachsignals $\hat{S}$ nicht negativ wird [13].

$$\hat{S}(\mu,\lambda) = |\hat{S}(\mu,\lambda)|e^{jarg(\hat{S}(\mu,\lambda)} \tag{2.18}$$

$$\hat{S}(\mu,\lambda) = |H(\mu,\lambda)||X|e^{jarg(Y(\mu,\lambda))} \tag{2.19}$$

2.5.2 Schätzung des Störabstands

Der „Decision-Directed“Ansatz stellt einen Kompromiss zwischen Störgeräuschreduktion und Sprachverzerrung dar [7]. Der geschätzte Störabstand $\hat{\eta}(\mu,\lambda)$ wird mittels eines IIR [7] Filters über zwei Rahmen geglättet. Der Störabstand (auch: apriori Signal zu Rauschverhältnis) des vorherigen Rahmens bestimmt auf Grund der hohen Werte für den Gewichtungsfaktor ($0.9 < \alpha_{DD} < 0.99$) maßgeblich den Wert des Störabstands des aktuell betrachteten Rahmens. Ein Wert nahe 1 für α_{DD} gewichtet das aposteriori Signal $\hat{\gamma}(\lambda,\mu)$ zu Rauschverhältnis des aktuellen Rahmens λ stärker. Hingegen führt eine Verringerung des Wertes zu einer stärkeren Glättung des Störabstands. Die Schätzung ist stabiler, jedoch werden schnelle Änderungen weniger wahrgenommen. Eigene Tests haben jedoch ergeben, dass der Effekt vernachlässigbar ist.

$$\hat{\eta}(\mu,\lambda) = \alpha_{DD} \cdot \hat{\eta}(\mu,\lambda-1) + (1-\alpha_{DD}) \cdot max(\hat{\gamma}(\mu,\lambda)-1,0) \tag{2.20}$$

[7] infinite impulse response, deutsch: unendlich lange Impulsantwort

Dabei sind der geschätzte apriori Störbabstand $\hat{\eta}(\mu, \lambda)$ und aposteriori Störabstand $\hat{\gamma}(\mu, \lambda)$:

$$\hat{\eta}(\mu, \lambda) = \frac{|\hat{S}(\mu, \lambda)|^2}{|\hat{N}(\mu, \lambda)|^2} \tag{2.21}$$

$$\hat{\gamma}(\mu, \lambda) = \frac{|X(\mu, \lambda)|^2}{|\hat{N}(\mu, \lambda)|^2} \tag{2.22}$$

Beim Decision Directed Ansatz basiert die Schätzung der Amplitude des Sprachsignals $\hat{S}(\mu, \lambda)$ auf der spektralen Gewichtung des verrauschten Signals des Rahmens λ unter der Verwendung des gewichteten apriori Störabstands $\hat{\eta}(\lambda-1, \mu)$ des vorhergehenden Rahmens 2.21. Dieser wiederum berechnet sich unter Verwendung der geschätzten Amplitude des Sprachsignals $\hat{S}(\mu, \lambda - 1)$. Es müssen aufgrund der Rekursion also Startwerte festgelegt werden. Diese werden hier zu Null gesetzt.

$$\hat{\eta}(\mu,\lambda) = \max(\hat{\eta}(\mu,\lambda), \eta_{min}(\mu,\lambda)); \tag{2.23}$$

Der minimale Störabstand ist auf $\eta_{min}(\mu, \lambda) = 0.01$, dies entspricht $-20\,\text{dB}$ begrenzt. Zu beachten ist, dass die oben genannten Berechnungen bis auf das verrauschte Signal geschätzte Signale verwenden.

2.5.2.1 Störgeräuschschätzung

Die Störgeräuschschätzung erfolgt nach dem Verfahren von [9], welches von der Wahrscheinlichkeit der Anwesenheit von Sprache (speech presence probability: kurz SPP) Gebrauch macht. Vorteile dieses sogenannten „SPP Trackers“ liegen in der geringeren Überschätzung des Rauschanteils.

2.6 Bewertung der konventionellen Störgeräuschreduktion (Wiener Filter)

Da die konventionelle Störgeräuschreduktion mit der Wiener Filterregel als Referenz für alle hier betrachteten psychoakustisch motivierten Verfahren verwendet wird, sollen im Folgenden kurz die Leistungsfähigkeit anhand instrumenteller Maße und typische Probleme beleuchtet werden. Die Filterung wird auf das Audiobeispiel mit dem Satz „Oak is strong“, welches einen 4 s langen Ausschnitt aus der Audiodatei ’/share/sounds/databases/TSPspeech/48k/track02.wav’ der Datenbank des IND darstellt, angewendet. Das gestörte Signal enthält weißes Rauschen (ggf. statt weißem Rauschen Bohrhammergeräusch) mit einem Störabstand von 5 dB.

2.6.1 Probleme bei der konventionellen Störgeräuschreduktion

Bei der konventionellen Störgeräuschreduktion treten Sprachverzerrungen auf. Neben einer zu starken Sprachdämpfung [8] bestimmter spektraler Anteile, im Besonderen der höheren

[8] gemeint ist die Dämpfung des klaren Sprachsignals, welches im verrauschten Signal enthalten ist

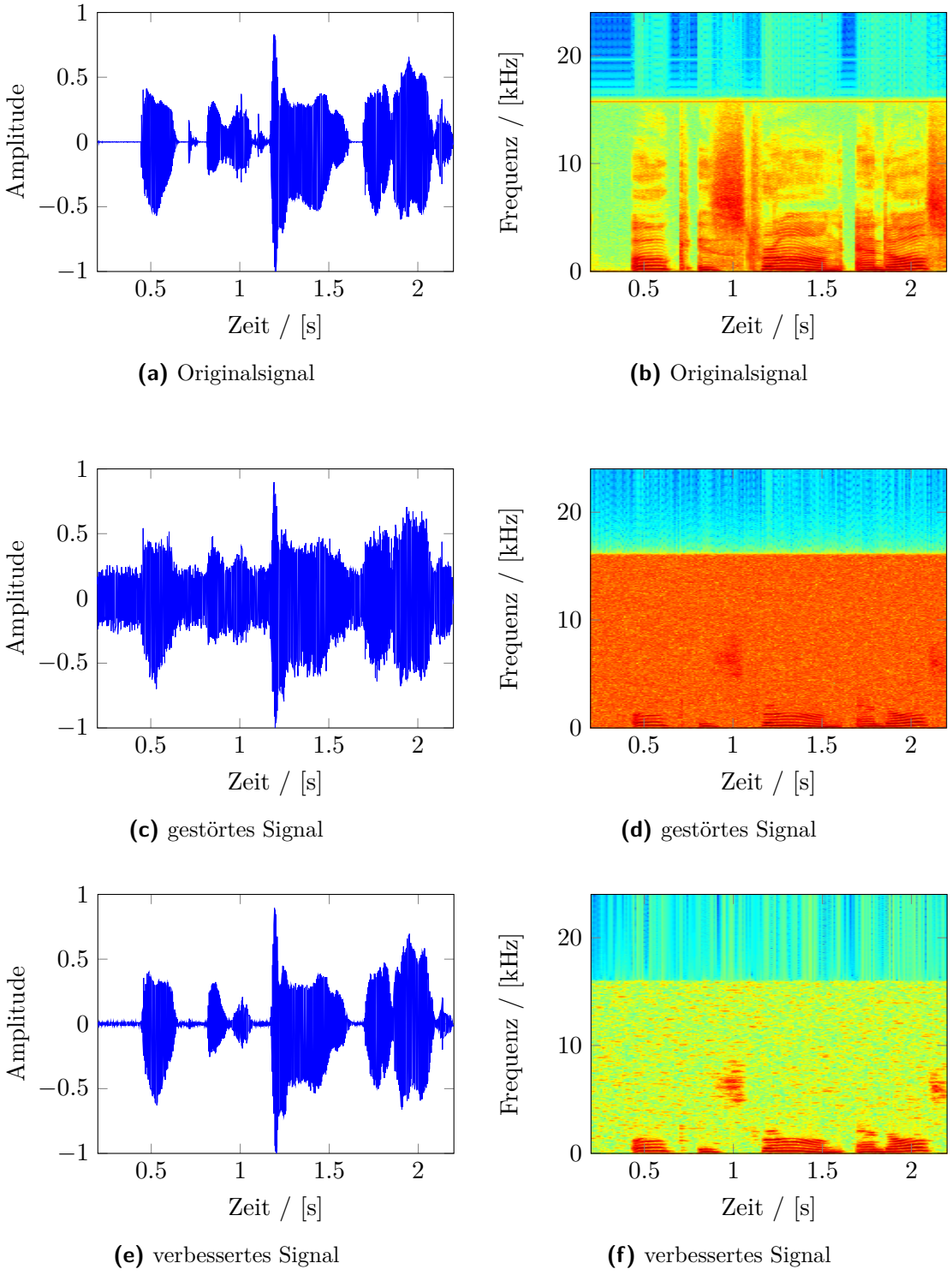
Frequenzanteile, welche zwar wie oben beschrieben einen geringen Anteil der Gesamtenergie ausmachen, dennoch aber das empfundene Klangbild merklich beeinflussen, wirkt besonders das sogenannte „Musical Noise“ störend. Als „Musical Noise“ bezeichnet man Artefakte, welche durch Spitzen im verbleibenden Rauschspektrum (Reststörung) nach der Filterung auftreten, [13]. Im Zeitbereich entsprechen diese Spitzen jeweils Tönen kurzer Dauer.

In den unteren Abbildungen 2.2 sind die Zeitverläufe und Spektogramme des originalen klaren Sprachsignals, des mit weißem Rauschen und einem Störabstand von 5 dB behafteteten verrauschten Signals und des nach Filterung mit der Wiener Regel verbesserten Signals dargestellt. Ein Beispiel für Sprachdämpfung kann man beim Vergleich der Zeitverläufe des Original- und verbesserten Signals in den Abbildungen 2.2a, 2.2e im Zeitbereich von ca. 0,7 bis 1,2 Sekunden sehen. Manche Ausschläge der Amplitude sind nicht mehr im verbesserten Signal vorhanden. Auffällig ist auch, dass bei sehr hohen Amplituden der Signalverlauf ähnlicher ist und fehlende Anteile vor allem bei niedrigen Amplituden des Sprachanteils auftreten. Mit sinkendem momentanem Störabstand nimmt die Rekonstruktionsfähigkeit des Filters also ab.
Während stark gedämpfte hochfrequente Anteile die Sprache „dumpf“ erklingen lassen, tritt das Musical Noise als störendes Geräusch im gesamten Frequenzbereich auf, vorwiegend und praktisch nur wahrnehmbar, falls der Störabstand sehr niedrig ist. Da die spektrale Gewichtung unter Verwendung des Störabstands auf der Schätzung des Rauschens basiert und diese Schätzung zu niedrigen Werten des realen (nicht bekannten) Störabstands immer schlechter wird, pflanzt sich der Fehler bei der Berechnung des geschätzten Störabstands fort, siehe Gleichungen 2.20, 2.21, 2.22. Der Fehler bei der Rauschspektralschätzung kann sich im resultierenden Filtergewicht vervielfacht haben:
Die Sprachschätzung wird beim Decision Directed Ansatz auf Basis des Störabstands des vorhergehenden Rahmens und der daraus resultierenden Gewichtung des gestörten Sprachsignals des vorhergehenden Rahmens berechnet. Ist also die Schätzung des Rauschens fehlerhaft, dann ist es auch die Schätzung des Sprachsignals des vorhergehenden Rahmens, und damit auch des Störabstands des aktuellen Rahmens. Der Fehler wird zwar aufgrund der IIR Filterung in 2.20 mittels Einbeziehen des geschätzten aposteriori Störabstands γ des aktuellen Rahmens verringert, dennoch treten durch die fehlerhafte Gewichtung des gestörten Signals verbleibende einzelne hervorstechende Spitzen im Spektrum auf. Dies ist im Spektogramm in Abbildung 2.2f als vereinzelte rote „Flecken“ zu sehen, die im Originalsignal (Abbildung 2.2b) nicht vorhanden sind.

Es ist anzumerken, dass auch im Falle einer perfekten Schätzung des Rauschens, also der Rauschanteil exakt bekannt ist, eine perfekte Rekonstruktion des Sprachsignals nicht möglich ist, da die Phasen des reinen Sprachsignals und des Rauschens unterschiedlich und unbekannt sind [13].

2.6.2 Typische Kennwerte

Im Folgenden sollen zum Vergleich mit den behandelten auf Psychoakustik basierenden Störgeräuschreduktionsverfahren tabellarisch die oben beschriebenen instrumentellen Maße präsentiert werden. Die Werte werden für ein 4s langes Signal berechnet. Ausschnitte dieses Signals sind in den den oberen Abbildungen (2.2a bis 2.2f) gezeigt. Die Abtastrate beträgt $f_s = 48\,\text{kHz}$. Vor der spektrale Gewichtung wird bei der Analyse ein Hannfenster



(a) Originalsignal **(b)** Originalsignal

(c) gestörtes Signal **(d)** gestörtes Signal

(e) verbessertes Signal **(f)** verbessertes Signal

Abbildung 2.2: Zeitsignal und Spektogramm des normierten Sprachsignals (a,b), des gestörten Signals (c,d), und des verbesserten Signals (e,f)

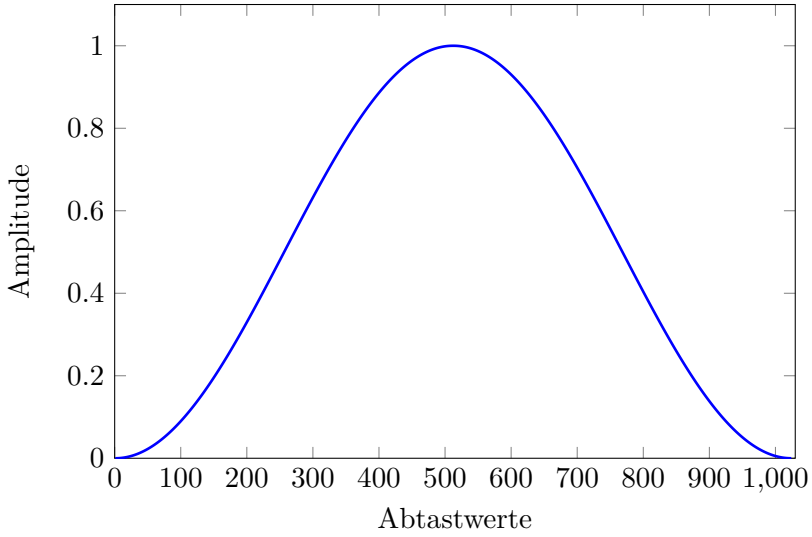


Abbildung 2.3: Hannfenster, welches auf Segmente des diskreten Signals im Zeitbereich zur Fensterung benutzt wird

in Abbildung 2.3 zur Segmentierung verwendet, um Leckeffekte zu vermeiden. Der Leckeffekt beschreibt das sich aus dem kurzen Beobachtungszeitraum durch die Segmentierung des Zeitsignals bei Anwendung der Fourierrtansformation ergebende „Leck“ der Spektrallinie zum Bsp. eines Sinustons hinzu benachbarten Spektrallinien. Das Spektrum ist dann verschmiert und weist neben der Spektrallinie für den Sinuston auch Anteile mit jedoch geringerer Spektralamplitude für benachbarte Frequenzen auf. Dieses Leck träte bei theoretisch unendlich langem Beobachtungszeitraum nicht auf - das Spektrum für den Sinuston besteht dann nur aus der einzelnen Spektrallinie. Das Hannfenster weist verglichen mit anderen Fenstern einen stärkeren Abfall bei hohen Frequenzen auf Kosten höherer Spitzen in den Nebenzipfeln auf [23].

Bei der Synthese kommt ein Rechteckfenster zum Einsatz. Konventionell wird das Hannfenster auf die Analyse und die Synthese aufgeteilt, was dann zu einem Fenster führt, welches sich aus der Wurzel des Hannfensters ergibt. Diese Methode ist zum Vergleichen mit den psychoakustischen Modellen weniger geeignet, da diese bei der Analyse des aus der konventionellen Störgeräuschreduktion gewonnenen verbesserten Signals die im Standard festgelegte Hannfensterung verwenden. Die Länge der segmentell angewendeten Fouriertransformation (DFT) beträgt $N = 1024$ oder $N = 2048$ Abtastwerte. Die zwei Längen werden jeweils in den mit der klassischen Störgeräuschreduktion zu vergleichenden psychoakustischen Störgeräuschreduktionsverfahren verwendet.
Für das sogenannte „overlap add“ Verfahren [23], welches die zyklische Faltungsoperation [9] in eine aperiodische Faltung überführt, wird der Fensterlänge auf die Hälfte der Länge der Fouriertransformation festgelegt und der Rest der Abtastwerte mit Nullen aufgefüllt („zeropadding“ genannt [23]). Die Überlappung von 50 % wird erreicht, indem die von dem Zeitsignal genommenen Fenster um die Hälfte ihrer Länge verschoben werden. Der Rahmenvorschub beträgt also bei einer DFT mit einer Länge von 2048 Punkten 512 Punkte.

[9] da die Fouriertransformation aufgrund der segmentellen Verarbeitung auf ein zeitlich begrenztes und damit nicht periodisches Signal angewendet wird

Maß | DFT Länge N	1024	2048
CD	25,37	25,50 dB
$SegSA$	4,89	4,88 dB
$SegNA$	22,19	22,23 dB
$SegDA$	17,29	17,35 dB
$SegSNR$	5,28	5,16 dB
$SegSpSNR$	7,56	7,52 dB
PESQ	1,4660	1,5770
STOI	0,866	0,887

Tabelle 2.1: Referenzkennwerte der konventionellen Störgeräuschreduktion mit Wiener Filter für zwei unterschiedlich lange Fouriertransformationen

In der Tabelle 2.1 sind Kennwerte der klassischen Störgeräauschreduktion unter Verwendung der Wiener Filterregel für die beiden verschieden Transformationslängen gegeben. Die Ergebnisse sind für alle instumentellen Maße nahezu gleich. Je nach den Ergebnissen der psychoakustischen Modelle reicht es möglicherweise aus, nur eine Länge zum Vergleich heran zu ziehen. Die Sprachdämpfung ca. 5 dB der konventionellen Störgeräuschreduktion ist deutlich niedriger als die Dämpfung des Störgeräuschs (ca. 22 dB). Damit ist die Differenz, welche maximiert werden soll relativ groß.

Alle psychoakustisch basierten Verfahren werden mit diesen Ergebnissen bewertet.

2.6.2.1 Probleme

Die Kennzahlen bilden aber nicht das Auftreten des Musical Noise ab, für das es bislang kein instrumentelles Maß gibt. Die Nachteile des Wiener Filters liegen also in der Empfindlichkeit gegenüber fehlerhafter Störgeräuschschätzung, welche das Auftreten von Musical Noise erhöhen. Allgemein eignet sich der in der konventionellen Störgeräuschreduktion zur leichten Milderung dieses Phänomens eingesetzte Decision Directed Ansatz wie auch die Wiener Filterregel selbst nur für stationäre Signale. Die Stationärität ist aber möglicherweise aufgrund der Fehlschätzung nicht mehr gegeben.

Die Fragestellung, ob die Sprachdämpfung möglicherweise und auch das Auftreten von Artefakten (vor allem Musical Noise) durch Einbeziehen der psychoakustischen Eigenschaften des menschlichen Gehörs reduziert werden kann, ist Thema dieser Arbeit.

2.7 Motivation zur psychoakustischen Störgeräuschreduktion

In der Abbildung 2.4 unten wird das allgemeine Problem bei Anwendung von spektraler Gewichtung mit dem Filtergewicht H veranschaulicht. Aufgetragen sind die Störleistungsdichte des Rauschens R_{qnqn} (blaue Kennlinie) und der Sprache R_{qsqs} (schwarze Kennlinie), sowie die Summe der beiden (rote Kennlinie), d.h. die gesamte Störleistungsdichte Rqq.

Die Störleistungsdichten stellen jeweils die Differenz zwischen nach der Filterung gewünschter verbleibender Leistungsdichte und tatsächlich verbleibender Leistungsdichte dar (Gleichungen 2.25). Sie sind jeweils antiproportional zu den jeweiligen Dämpfungen. Alle

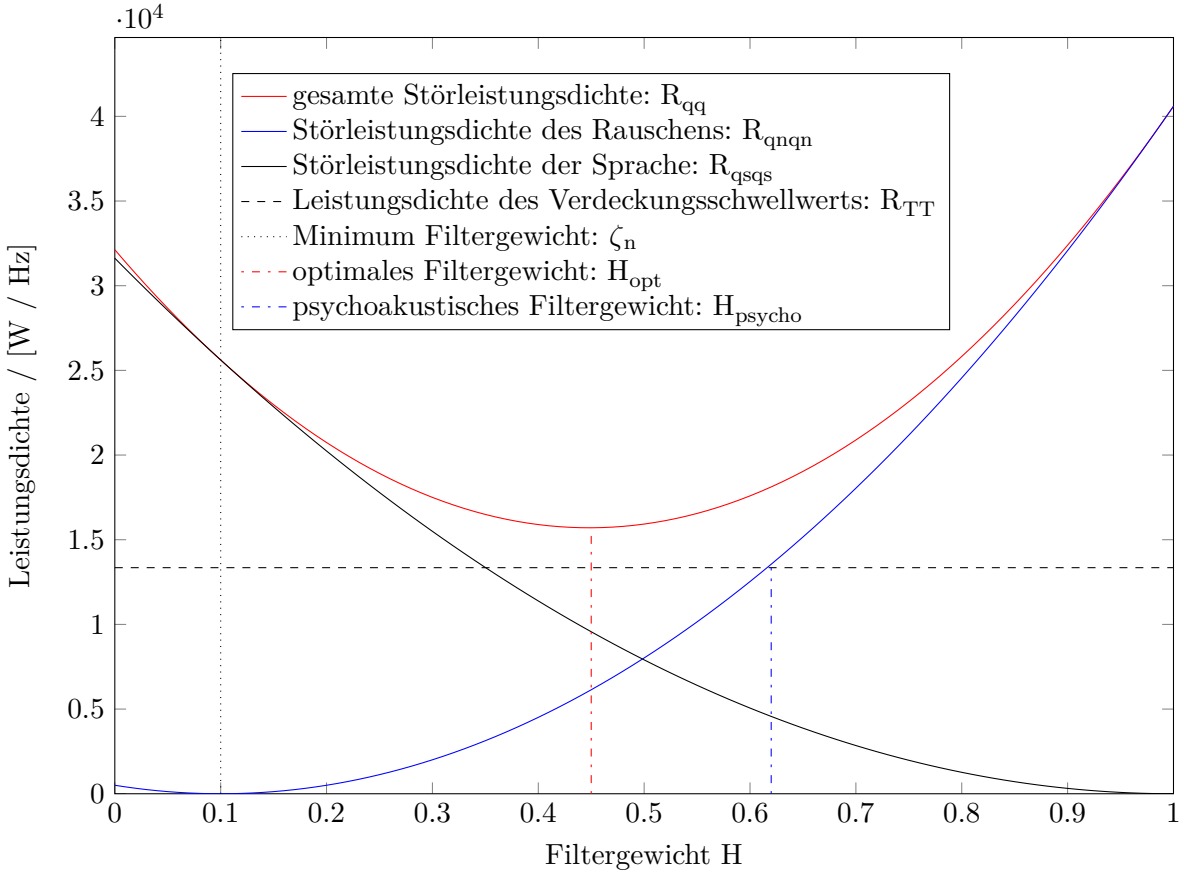


Abbildung 2.4: Filtergewicht in Abhängigkeit der Sprach- und Rauschdämpfung

Größen werden hier für eine normierte Frequenz Ω zu einem bestimmten Zeitpunkt betrachtet. In der praktischen Umsetzung entsprächen die Größen ihren von Frequenzlinie μ und Rahmen λ abhängigen Werten.
Ein verrauschtes Signal wird zur Filterung mit dem Filtergewicht H, das Werte von null bis eins annehmen kann, entsprechend 2.3 gewichtet. Im ersten Fall ($H = 0$), wird das verbesserte Signal null: es enthält weder einen Rauschanteil, noch einen Sprachanteil. Die Störleistungsdichte des Rauschens R_{qnqn} verschwindet, die Störleistungsdichte der Sprache R_{qsqs} ist maximal.

Beträgt der Wert des Filtergewichts eins, wird der Sprachanteil nicht verzerrt (die Störleistungsdichte R_{qsqs} der Sprache verschwindet), allerdings ist dann auch die Störleistungsdichte des Rauschens maximal. Die Gleichungen unten verdeutlichen diesen Zusammenhang:

$$R_{qsqs}(\Omega) = [1 - H(\Omega)]^2 R_{ss}(\Omega) \tag{2.24}$$

$$R_{qnqn}(\Omega) = [\zeta_n - H(\Omega)]^2 R_{nn}(\Omega) \tag{2.25}$$

ζ_n dient zum Einstellen des verbleibenden Rauschens, in dem das Filtergewicht nach unten auf diesen Wert begrenzt wird. Zur Verringerung des Rauschens nimmt das Filtergewicht H je nach Störabstand nun einen Wert zwischen null und eins an, welcher

die gesamte Störleistungsdichte R_{qq} (rot) minimiert. Dies entspricht der Minimierung des Fehlersignals $Q(\Omega)$, welches sich aus der Differenz zwischen gewünschtem Ausgangssignal $\check{S}(\Omega)$ und $\hat{S}(\Omega)$ ergibt.

Ziel: Verdeckung der Störung des gewünschten Rauschens R_{qnqn} (Leistungsdifferenz zwischen gewünschtem und tätsächlichem Rauschen). Die Differenz zwischen gewünschtem $\check{S}(\Omega)$ und tatsächlichem Ausgangssignal $\hat{S}(\Omega)$

$$Q(\Omega) = \check{S}(\Omega) - \hat{S}(\Omega) \tag{2.26}$$

beschreibt das Fehlersignal. Bei der konventionellen Störgeräuschreduktion wird das Fehlersignal minimiert, so dass die gesamte Störleistungsdichte (Erwartungswert des Quadrats des Fehlersignals) minimal wird:

$$R_{qq} = E[Q(\Omega)^2] \tag{2.27}$$

Der Wert ist als senkrecht gestrichelte Linie in der Abbildung 2.4 zu sehen und wird als optimales Filtergewicht H_{opt} bezeichnet. Dieses Filtergewicht kommt bei der konventionellen Störgeräuschreduktion zur Anwendung und führt zu den genannten Sprachverzerrungen und Artefakten.

2.7.0.2 Ausnutzung von Verdeckungseffekten

Es ist allgemein bekannt, dass Geräusche andere Geräusche verdecken (maskieren). Redet eine Person in einem Raum relativ laut (relativ hoher Schalldruckpegel), wird das Rauschen z.B. der Klimaanlage oder im Allgemeinen ein Geräusch mit geringerem Schalldruckpegel leiser bis gar nicht mehr wahrgenommen. Auch die Erholungsphase des menschlichen Gehörs nach einem lauten Ereignis reduziert die Wahrnehmung von darauffolgenden Ereignissen, sofern diese einen bestimmten Schalldruckpegel nicht übersteigen. Das erste Phänomen wird als spektrale Verdeckung [10], das zweite als temporale Verdeckung bezeichnet. In der Abbildung 2.4 ist der Verdeckungsschwellwert R_{TT} (gestrichelte schwarze Linie) für die betrachtete Frequenz Ω und einen Zeitpunkt repräsentiert. Der auf Basis der Leistungsdichte der Sprache berechnete Schwellwert gibt an, bis zu welcher Leistungsdichte der Störung R_{qnqn} verdeckt wird, also nicht „hörbar“ macht. Solange die Störleistungsdichte also unterhalb des Schwellwerts bleibt, kann das Filtergewicht H gegenüber H_{opt} erhöht werden, ohne dass mehr wahrnehmbares Rauschen im verbesserten Signal auftritt. Im verbesserten Signal ist dann gegenüber der Filterung mittels konventioneller Störgeräuschereduktion (also H_{opt}) die Sprachdämpfung geringer (bzw. die Störleistungsdichte der Sprache R_{qsqs}) und zumindest theoretisch der Höreindruck des verbleibenden Rauschens gleich, da idealerweise die Verdeckung die Differenz des höheren Rauschanteils und des nach Verarbeitung mit der konventionellen Störgeräuschreduktion erzielten Rauschanteils perfekt maskiert.

Zu beachten ist, dass es sich bei dieser theoretischen Betrachtung bei den Leistungsdichten nicht um Schätzwerte, sondern um die tatsächlichen Werte handelt.

Die Idee der spektralen und zeitlichen Maskierung von Signalen zielt darauf ab, für das menschliche Gehör nicht hörbare Anteile auch möglichst wenig bis gar nicht zu filtern,

[10] auch „Maskierung“ - die Begriffe werden in der Literatur synonym verwendet

da dabei aufgrund ungenauer Schätzung Artefakte oder eine zu hohe Dämpfung auftreten können. Umgesetzt wurde dies bisher nur beim verlustbehafteten Codieren (Komprimieren) von Audiosignalen wie z.B mit dem MPEG1 Layer 3 Codec. Dabei wird berechnet in wie weit das zu kodierende Audiosignal durch das durch Quantisierung entstehende Rauschen verdeckt und somit variable Auflösungen je nach Hörbarkeit verwendet werden. Übersteigt das Rauschen den berechneten Schwellwert, muss mit einer höheren Auflösung quantisiert werden. Bei Anwendung von Verdeckungseffekten zur Störgeräuschreduktion ist das Störgeräusch schon vorhanden und nicht beeinflussbar, d.h. das Rauschen wird nicht vollständig, sondern nur teilweise, schlimmstenfalls gar nicht verdeckt. Allgemein kann sowohl Rauschen das Nutzsignal (die Sprache) verdecken, als auch umgekehrt. Häufig ist der durchschnittliche Störabstand so groß, dass in den Sprechzeiten die Sprache im Zeitverlauf größtenteils die im Vergleich zum Rauschen höhere Amplitude aufweist. Die in dieser Arbeit verwendeten Filterregeln (Wiener und angepasste Filterregeln, Filter nach [13]) tragen dem Phänomen, dass Rauschen Sprache maskieren kann, mittels für diesen Fall sehr kleinem Filtergewicht Rechnung. Das verrauschte Signal und damit auch das Rauschsignal werden dann weitestgehend unterdrückt.

2.7.1 Verdeckungsbasierter Filter

Gustafsson hat in seiner Arbeit [13] eine Filterregel H_{IND} vorgestellt, die das verrauschte Signal psychoakustisch gewichtet: Ist die Leistungsdichte des Schwellwerts der geschätzten Sprache $R_{TT}(\Omega)$ gegenüber der Leistungsdichte des geschätzten Rauschens groß, so ergibt sich ein großes Filtergewicht und eine geringe Sprachdämpfung bei gleichzeitiger im Sinne der Hörbarkeit minimaler Störgeräuschreduktion [11] , zumindest unter der Annahme, dass die Schätzung der Rauschleistungsdichte und der darauf basierenden Berechnung der Schwellwerte nahezu den tatsächlichen Größen entspricht.

$$H_{IND}(\Omega) = min\left(\sqrt{\frac{\hat{R}_{TT}(\Omega)}{\hat{R}_{nn}(\Omega)}} + \zeta_n, 1\right) \tag{2.28}$$

Umgekehrt führt z.B. im Fall von Abwesenheit von Sprache oder sehr niedrigen Störabständen die Berechnung zu einem Filtergewicht nahe dem Wert Null bzw. ζ_n. Das heißt verdeckt die Sprache das Rauschen nicht oder nicht ausreichend, so wird das verrauschte Signal stark gedämpft und erfährt damit eine Störgeräuschreduktion.

Für den Dämpfungsfaktor ζ_n wurde $20\log(\zeta_n) = -15\,\text{dB}$ gewählt, [13]. Nur bei Signal-zu-Rauschverhältnissen unter 0 dB sollte ζ_n größer gewählt werden, um mögliche Störungen in der Sprache zu vermeiden [13].

2.7.2 Zweistufige Störgeräuschreduktion

Zur Berechnung der Schwellwerte muss das Sprachsignal geschätzt werden. Dies geschieht mittels der konventionellen Störgeräuschreduktion [12]. Die Struktur des Gesamtsystems

[11] eine stärkere Dämpfung des Störgeräuschs würde aufgrund der Verdeckung idealerweise zum gleichen Höreindruck *bzgl. des Rauschens* führen

[12] in dieser Arbeit mittels Wiener Filter

ist als zweistufige Störgeräuschreduktion in 2.5 dargestellt. Die grün maskierten Blöcke repräsentieren die erste Stufe, welche aus der konventionllen Störgeräuschreduktion besteht. Die rot umrahmten Blöcke stellen die Erweiterung des Systems durch die zweite, der psychoakustischen Stufe mit dem Filtergewicht H_{psycho} dar. Das zeitdiskrete Signal $x(k)$ wird segmentiert, mittels Hannfenster gefenstert und dann einer Fouriertransformation unterzogen. Es folgt die Schätzung der Leistungsdichte des Rauschens $\hat{R}_{nn}(\Omega)$ wie im Abschnitt zur konventionellen Störgeräuschreduktion beschrieben. Die Schätzung des Spektrums der Sprache erfolgt durch die erste Stufe, welches zu dem Musical Noise behafteten verbesserten Signal, wie in Abbildung 2.2e und 2.2f gezeigt, führt. Die in dieser Arbeit verwendeten psychoakustischen Modelle berechnen die geschätzte Leistungsdichte des Schwellwerts $R_{TT}(\Omega)$ basierend auf diesem Spektrum. Das Filtergewicht H_{psycho} resultiert aus den beiden Leistungsdichten $\hat{R}_{nn}(\Omega)$ und $\hat{R}_{TT}(\Omega)$. Das oben vorgestellte Filtergewicht nach Gustafsson stellt eine mögliche Alternative für die Filterung in der zweiten Stufe dar. In dieser Arbeit werden auch andere Gewichtungen untersucht, siehe Kapitel 5. Anschließend wird das verrauschte Signal $X(\Omega)$ im Frequenzbereich mit dem Filtergewicht multipliziert und das geschätzte Sprachsignal $\hat{S}_2(\Omega)$ mittels dem overlap add Verfahren [23] und inverser Fouriertransformation in den Zeitbereich zurück transformiert. Erwartet wird ein verbessertes Signal $s(k)$, welches weniger Musical Noise und wümschenswerterweise eine geringere Sprachdämpfung gegenüber dem oben vorgestellten Referenzsystem aufweist.

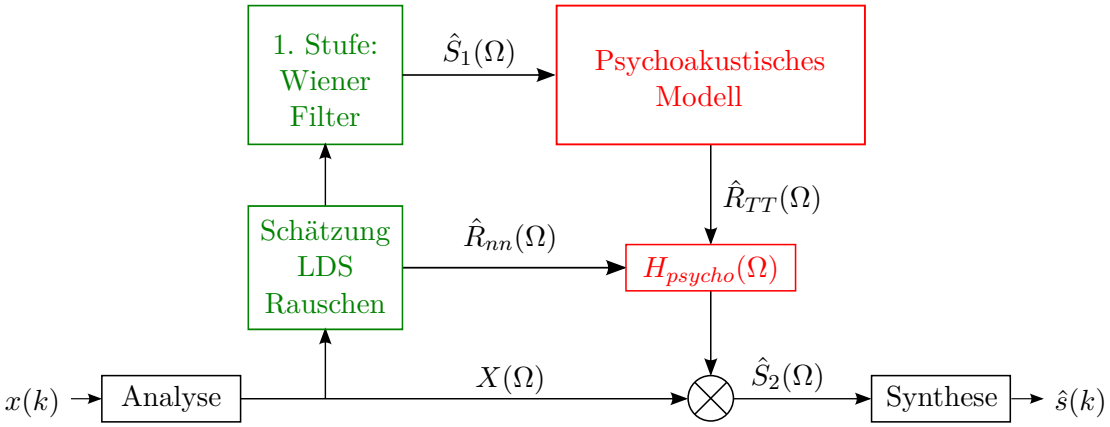


Abbildung 2.5: Zweistufige Störgeräuschreduktion bestehend aus konventioneller Störgeräuschreduktion (erster Stufe: Wiener Filter - grün) und zweiter Stufe mit psychoakustischem Modell

2.7.3 Mögliche Grenzen psychoakustischer Störgeräuschreduktion

Da der Fehler der Schätzung der Sprache durch die erste Stufe sich bei der Berechnung der Schwellwerte / Anregungsmuster fortpflanzt könnte man erwarten, dass das Ergebnis der zweiten Stufe, sich gegenüber der ersten Stufe möglicherweise noch verschlechtert. Die Genauigkeit der Schwellwerte ist in jedem Fall durch die Genauigkeit der Schätzung der Sprache begrenzt und damit möglicherweise auch die Leistungsfähigkeit der zweistufigen Störgeräuschreduktion. Dies soll unter anderem in dieser Arbeit untersucht werden. Im

folgenden Kapitel wird auf die psychoakustische Grundlagen eingegangen.

Kapitel 3

Psychoakustik

Im Kapitel Grundlagen wird die zweistufige Störgeräuschreduktion beschrieben, die als zweite Stufe ein psychoakustisches Modell verwendet, um die Verdeckungsschwellwerte für das nachgeschaltete Filter H_{psycho} zu berechnen. Bevor die Modelle zur Berechnung des oben erwähnten Verdeckungsschwellwerts zur Ausnutzung psychoakustischer Effekte im nächsten Kapitel vorgestellt werden, behandelt dieses Kapitel die größtenteils in den Modellen repräsentierten wichtigsten psychoakustischen Phänomene und Eigenschaften des menschlichen Gehörs.

3.1 Das Gehör

Die Verarbeitung von Audiosignalen durch das menschliche Gehör findet makroskopisch gesehen auf zwei Ebenen statt, die sich in der Signalart verdeutlichen: Schall (Gehörgang) und elektrisches Signal (neuronale Ebene).
In Abbildung 3.1 ist der Signalweg durch das menschliche Gehör als Blockdiagramm dargestellt. Während im Außen-, Mittelohr und auch noch im Innenohr das Signal in Form von Schall vorliegt, wird es im Innenohr durch die inneren Haarzellen in ein elektrisches Signal gewandelt und dann auf neuronaler Ebene im Gehirn weiterverarbeitet. Für eine detaillierte Darstellung des Ohres sei auf [3] verwiesen, eine kurze Einführung findet sich in [12].

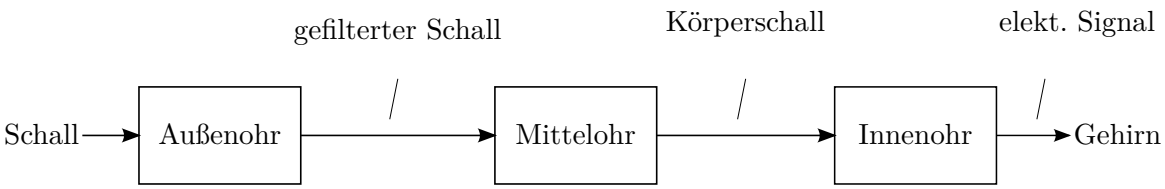


Abbildung 3.1: Das menschliche Gehör als System

3.1.1 Außenohr und Mittelohr

Der Luftschall erreicht das Außenohr und erfährt eine Bandpassfilterung. Das aus Hammer, Amboss und Steigbügel bestehende Mittelohr übertragt den im Außenohr noch sich in Luft ausbreitenden Schall, welcher auf das Trommelfell trifft, als Körperschall auf das Innenohr. Dort breitet sich der Schall in der Innenohrflüssigkeit in Form von Wanderwellen aus. Die Übertragungsfunktion $A(f)$ des Außen- und Mittelohres ist unten in Abbildung 3.2[1] zu sehen.

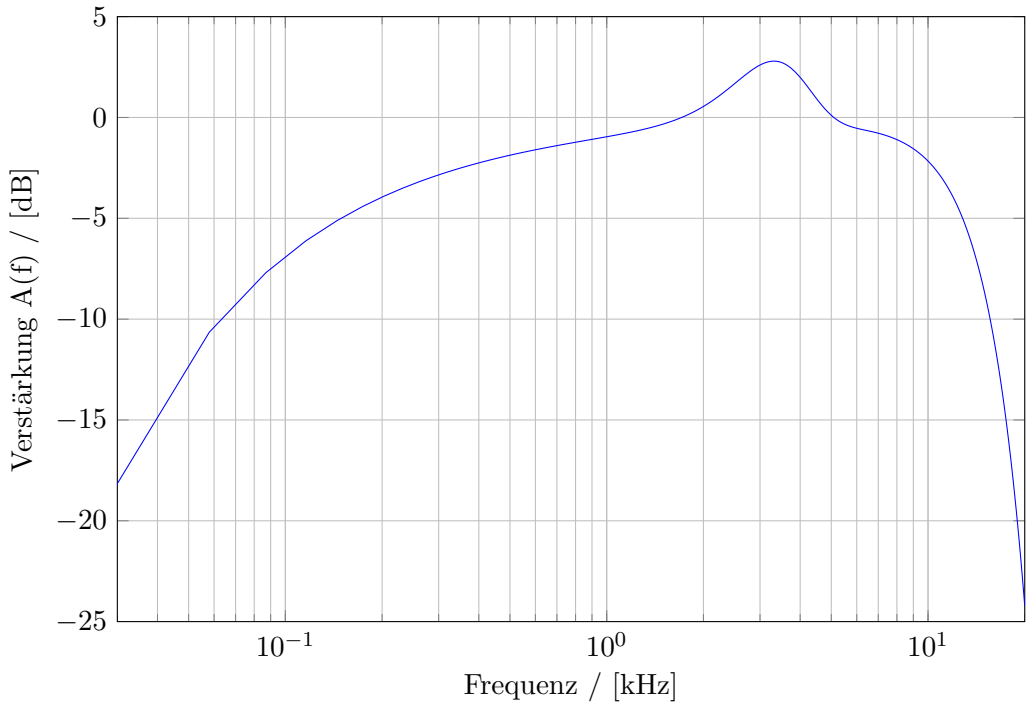


Abbildung 3.2: Übertragungsfunktion des Außen- und Mittelohres

Frequenzen unterhalb von 100 Hz werden stark gedämpft, darüber liegende Frequenzen bis ungefähr 1 kHz schwächer. Im Frequenzbereich von 1,8 kHz bis 5 kHz findet keine *aktive* sondern passive Verstärkung statt. Das Außen- und Mittelohr stellen ein passives System dar. Die Verstärkung in diesem Bereich ist auf die Superposition am Außenohr der vom Torso zu diesem reflektierten Schallwellen zurückzuführen. Daher sind Audiosignale in diesem Frequenzbereich für den Menschen besonders gut wahrnehmbar.

3.1.2 Innenohr und neuronale Weiterverarbeitung

Das Innenohr ist wesentlicher Bestandteil der psychoakustischen Modelle, da dort die eigentliche Verarbeitung von Audiosignalen beginnt, die bzgl. der Komplexität die Filterung durch Außen- und Mittelohr deutlich übersteigt. Das Innenohr besteht im Wesentlichen aus der Gehörschnecke (Cochlea) in deren Inneren sich die Basilarmembran und darauf die inneren Haarzellen befinden. Die Haarzellen generieren elektrische Signale, sogenannte

[1] entsprechend der Formel aus [37]

„Aktionspotentiale“ (neuronale Aktivität) als Antwort auf die Vibration der Basilarmembran, welche durch den einfallenden Schall zu Querauslenkungen angeregt wird [37] und zur Ausbildung von Wanderwellen führt [14]. Die Basilarmembran eines ausgewachsenen Menschen ist ungefähr 32 mm lang [14]. Am Anfang der Cochlea, also an dem Ende der Basilarmembran, welches dem Außenohr näher ist, werden hohe Frequenzen verarbeitet und zum anderen Ende hin die niedrigeren Frequenzen. Dies ergibt sich aus der bei niedrigen Frequenzen größeren Wellenlänge der sich in der Cochlea ausbreitenden Wanderwellen bei niedrigen Frequenzen. Ist der ins Gehör eintretende Schall ein Sinuston, so ist die daraus im Innenohr entstehende Amplitude der Wanderwelle an der entsprechenden Resonanzstelle auf der Basilarmembran maximal. Es findet eine nicht-lineare Frequenz-Ort-Transformation [28] mit nachfolgender Umwandlung von Schall in elektrische Signale statt [37]. Das an dieser Stelle entstehende elektrische Signal ist verglichen zu dem Signal benachbarter Stellen analog zum Maximum der Wanderwelle maximal, sofern ein bestimmter Schalldruckpegel nicht überschritten wird. Durch nichtlineare Verzerrung und Intermodulationsprodukte mehrerer Frequenzen kann es zu weiteren Maxima auf der Basilarmembran kommen [6], [40].

3.1.2.1 Frequenzselektion

Die äußeren Haarzellen verstärken an der jeweiligen Resonanzstelle die Schwingungen durch anatomische Beeinflussung des Basilarmembran-Hörschnecken Verbunds. Die inneren Haarzellen erfahren dort dadurch eine stärkere Anregung. Hingegen steigt die Dämpfung der Schwingung mit zunehmender Entfernung an, sodass sich die Wanderwelle der jeweiligen Frequenz nicht (bzw. nur mit deutlich geringerer Amplitude) über ihre Resonanzstelle hinaus ausbreitet. Diese durch die äußeren Haarzellen repräsentierte rückkoppelnde und damit anpassende Verstärkung erhöht den Dynamikbereich des menschlichen Gehörs [12] und die Frequenzselektivität, vgl. dazu auch Abbildungen 3.9 und 3.11. Die Frequenzantwort der Basilarmembran ist also amplitudenabhängig.

Die Eigenschaften des Gehörs werden nicht nur durch das Gehörorgan selbst bestimmt, sondern auch durch periphere Bereiche des menschlichen Körpers. Die Strömung des Blutes der umliegenden Blutgefäße verursacht Körperschall, der als „Blutrauschen“ bezeichnet wird. Dieser wird genauso von den inneren Hörzellen auf der Basilarmembran detektiert wie der durch das Außenohr einfallende Schall, jedoch wird das Blutrauschen auf der nachfolgenden neuronalen Ebene teilweise unterdrückt. Die spontane Aktivität der Nerven selbst generiert Rauschen. Beide Rauschsignale werden als internes Rauschen IR zusammengefasst [37]: Das interne Rauschen wird durch folgende Gleichung beschrieben und in Abbildung 3.3 ist dessen Schalldruckpegel über der Frequenz aufgetragen.

Mathematisch ist das interne Raschen durch folgende Gleichung repräsentiert:

$$IR(f) = 0{,}4 \cdot f^{-0{,}8}\,, \text{ in dB} \tag{3.1}$$

Die Frequenz f wird dabei in kHz angegeben. Der Faktor 0.4 repräsentiert die oben genannte neuronale Unterdrückung des Blutrauschens. Besonders im unteren Frequenzbereich bis 100 Hz ist der durch das Blutrauschen verursachte Schalldruckpegel mit Werten bis 20 dB gegenüber dem in das Ohr einfallenden Schall nicht vernachlässigbar. Ein weiterer neuronaler Faktor findet sich in der Abbildung 3.2 dargestellten Übertragungsfunktion.

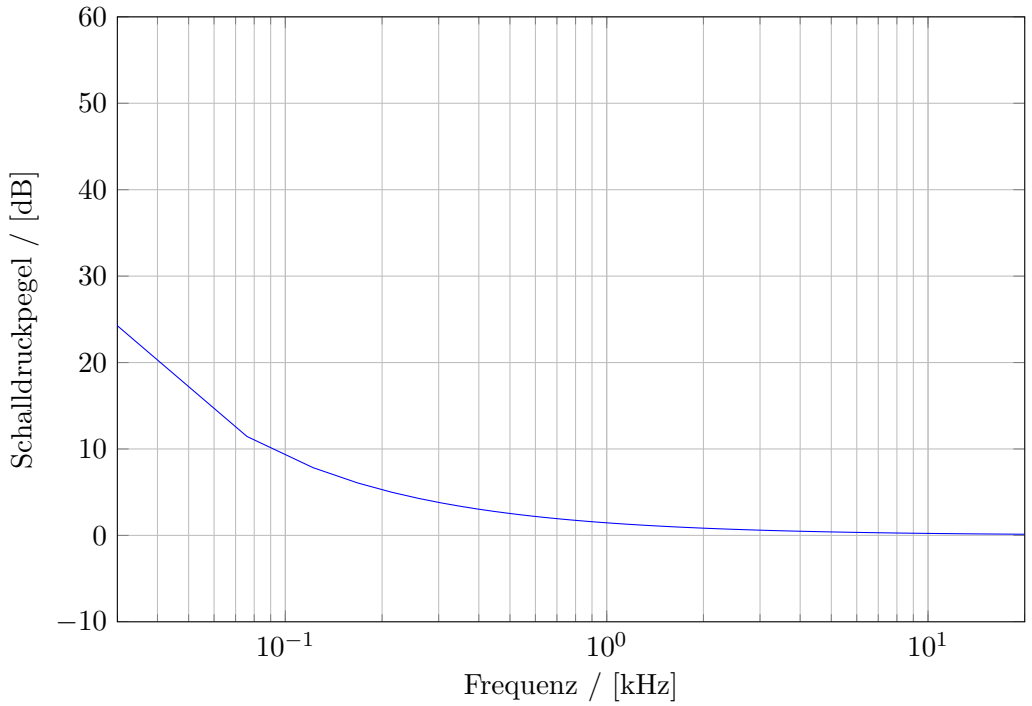


Abbildung 3.3: Schalldruckpegel des frequenzabhängigen internen Blutrauschens nach [37]

Dieser verstärkt die Dämpfung in den Randbereichen. Allerdings ist der neuronale Einfluss auf diese Funktion (Faktor −0,6) Gleichung 3.2) vernachlässigbar klein, weshalb in der Literatur nur von der Übertragungsfunktion des Außen - und Mittelohres gesprochen wird.

$$A(f) = -0{,}6 \cdot 3{,}64 \cdot f^{-0{,}8}\ ,\ \text{in [dB]} \tag{3.2}$$

Auch hier ist die Frequenz in kHz anzugeben. In der Gesamtheit führen die Eigenschaften von Außen-, Mittel- und Innenohr sowie der neuronalen Ebene zur absoluten Hörschwelle[2] nach [14] (blaue Linie in Abbildung 3.4) und den Kontouren gleicher gehörrichtiger Lautstärke, den sogenannten „Isophonen“ (Abbildung 3.11 in Abschnitt) 3.3. Dies lässt sich leicht nachvollziehen, wenn man die gespiegelte Übertragungsfunktion der Abbildung 3.2 mit der unteren Abbildung vergleicht.

Mathematisch lässt sich die absolute Hörschwelle $absHS$ in Abhängigkeit der Frequenz f, welche in kHz angegeben wird, mit der Gleichung 3.3 beschreiben.

$$absHS = 3{,}64 \cdot f^{-0{,}8} - 6{,}5e^{(-0{,}6\cdot(f-3.3)^2} + 10^{-}3 \cdot f^4\ ,\ \text{in [dB]} \tag{3.3}$$

Der Verstärkungsbereich von größer 0 dB in der Übertragungsfunktion findet sich in der Kurve der absoluten Hörschwelle im ungefähr gleichen Frequenzbereich wieder (zwischen gestrichelter schwarzer und magenta gefärbter Linie). Das absolute Minimum ist mit der roten Linie markiert. In diesem Bereich und darüber hinaus (grün markierter Bereich) sind typische Frequenzen 120 bis 7 kHz und Schalldrücke der Sprache anzusiedeln [14]. Geringe

[2] auch absoluter Schwellwert genannt

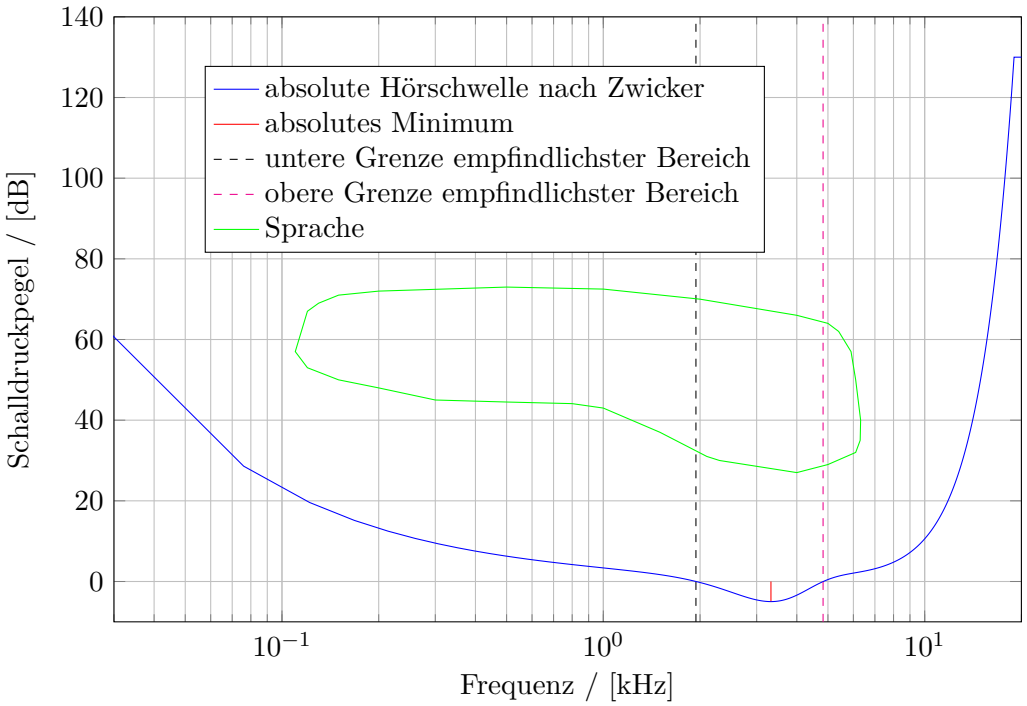


Abbildung 3.4: absolute Hörschwelle und Bereich des frequenzbezogenen Schalldruckpegels

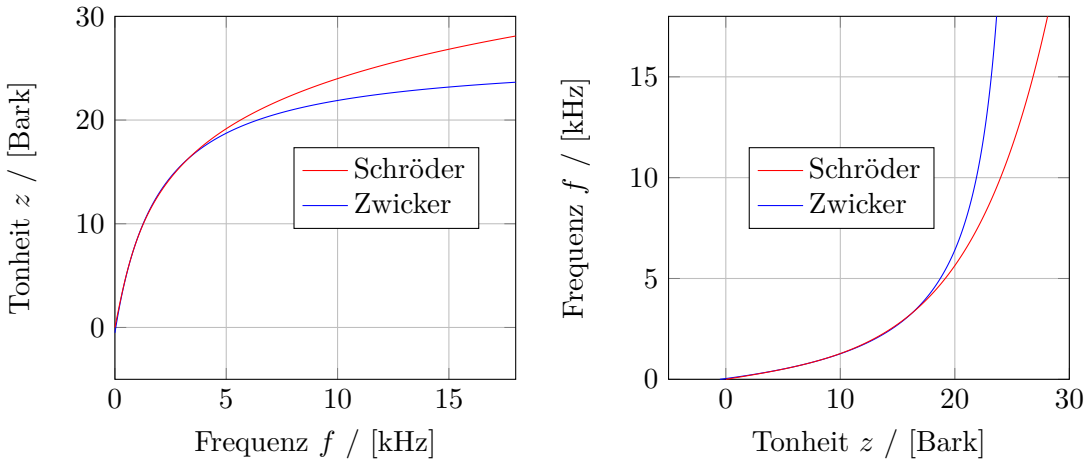
Anteile von Sprache finden sich darüber hinaus bis zu einer Frequenz von 14 kHz, [16]. Für Frequenzen größer 10 kHz und auch für Frequenzen kleiner 100 Hz steigt die Schwelle stark an. Die absolute Hörschwelle beschreibt die Schwelle des Schalldruckpegels, bei der die Versuchspersonen mit 50,% einen Ton wahrnehmen. Die Hörschwelle ist von Mensch zu Mensch individuell, liegt für junge Menschen mit geringer Varianz auf der blauen Linie der Abbildung 3.4 und nimmt mit dem Alter gerade bei hohen Frequenzbereichen ab 7 kHz zu. Die Grenzen des vom menschlichen Gehör wahrnehmbaren Frequenzbereichs liegen bei 16 Hz für die niedrigste Frequenz, während die Obergrenze von bis unter 10 kHz für ältere Menschen bis 20 kHz bei Kindern altersbedingt stark variiert [16]. In folgenden Unterabschnitten werden die Eigenschaften des Innenohres und der neuronalen Ebene weiter vertieft.

3.1.2.2 Spektrale Auflösung

Die nicht lineare Frequenz zu Ort Transformation führt zusammen mit der linearen Verteilung der inneren Haarzellen entlang der Basilarmembran zu einer nicht linearen Wahrnehmung der Frequenzen, dieses Phänomen wird „Tonheit“ (engl. pitch) genannt [37]. Das heißt, die subjektiv empfundene Tonhöhe unterscheidet sich von der physischen Tonhöhe, welche durch die Schallfrequenz gegeben ist [25]. Diese Tonhöhenwahrnehmung wird als „Tonheit“ bezeichnet. Es existieren einige Approximationen für die Beschreibung der Transformation vom Frequenzbereichs zur Tonheit. Am weitesten verbreitet ist die Transformation zu einer sogenannten „Barkskala“, von der verschiedene Versionen existieren: die Barkskala nach Zwicker [43] und dessen Korrekturen hoher Frequenzen durch Traun-

müller [39] (Gleichung 3.4 unten), sowie die Approximation nach [28] (Gleichung 3.5). Die Barkskala stellt eine mögliche Approximation der Bänder des Gehörs dar, bei denen die Bandbreite für niedrige Frequenzen niedrig und zu hohen Frequenzen hin immer größer werden [11].

Diese beiden Approximationen finden sich in den verwendeten psychoakustischen Modellen, welche in Kapitel 4 behandelt werden wieder. Darüber existieren weitere Transformationen, wie die sogenannte „Äquivalentrechteckbandbreite“ [3] und darauf basierende Modellierungen mittels Filterbänken, wie z.B die Gammatonfilterbank [33], der auditive Filter [27] oder wie der in Kapitel4 vorgestellten Filterbank. In den unteren Abbildungen 3.5 sieht man die Verläufe der Barkskalen nach Zwicker [43] und Schröder [28]. Während die Skalen bis zu einer Frequenz von $f = 5$ kHz nahezu identisch sind, weichen die Werte z für die Tonheit ab 8 kHz signifikant voneinander ab. Die Anwendung der unterschiedlichen Skalen sollte in relativ kleinen Unterschiede zwischen den verbesserten Sprachsignalen resultieren, da für diese wie oben beschrieben, nahezu die gesamte Energie unterhalb von Frequenzen von 7 kHz auftritt. Die Transformationen der verschiedenen Barkskalen sind



(a) Transformation Frequenz zu Tonheit **(b)** Rücktransformation Tonheit zu Frequenz

Abbildung 3.5: Approximationen der Barkskala

in den Gleichungen 3.4 und 3.5 gegeben. Dabei kennzeichnet der Index z die Transformationsvorschrift nach Zwicker und der Index s diejenige nach Schröder.

$$z_z = 13 \cdot \arctan\left(0{,}76 \cdot \frac{f}{1000\,\text{Hz}}\right) + 3{,}5 \cdot \arctan\left(\frac{f}{7500\,\text{Hz}}\right)^2 \text{, in [Bark]} \tag{3.4}$$

$$z_s = 7 * \ln\left(\frac{f}{650\,\text{Hz}} + \sqrt{1 + \left(\frac{f}{650\,\text{Hz}}\right)^2}\right) \text{, in [Bark]} \tag{3.5}$$

Die Rücktransformation von der Tonheit (also der Skala der empfundenen Tonhöhe) wird

[3] engl.: Equivalent Rectangular Bandwidth, kurz: ERB

analog mit den Gleichungen 3.6 und 3.7 mathematisch beschrieben.

$$f_z = 1000z \cdot \left(\frac{e^{0{,}219 \cdot \frac{z}{\text{Bark}}}}{352} + 0{,}1 \right) - 32 \cdot e^{-\frac{3}{20} \cdot (\frac{z}{\text{Bark}} - 5)^2} \tag{3.6}$$

$$f_s = 650 \cdot \sinh(\frac{z}{\text{Bark}}/7) \tag{3.7}$$

3.1.2.3 Frequenzgruppen und Bänder

Die Amplitude der Anregung der Basilarmembran durch einen Sinuston ist bei der Resonanzstelle maximal und neben der Resonanzstelle flach abklingend. Die oben erwähnte Frequenzselektion vgl. Abschnitt 3.1.2.1 verläuft also nicht „scharf“ entlang der Basilarmembran. Daraus ergibt sich eine kontinuierliche Transformation vom Frequenzbereich zur Tonheit. Ab einer gewissen Entfernung zur Resonanzstelle werden die Amplituden jedoch so klein, dass sie den absoluten Schwellwert wie in Abbildung 3.4 unterschreiten. Zwicker hat in [41] Ergebnisse seiner Untersuchungen vorgestellt, die eine Frequenzzuordnung durch das Innenohr in Frequenzgruppen stützen. Die Frequenzgruppen sind so definiert, dass ein Sinuston in der betrachteten Frequenzgruppe durch weitere Erhöhung der Bandbreite eines anfangs auf die Frequenzgruppenbandbreite begrenzten Rauschsignals über die Frequenzgruppe hinaus, nicht mehr zum Anstieg der Mithörschwelle führt. Als Mithörschwelle bezeichnet man den Schalldruckpegel, der benötigt wird, um einen Sinuston in Anwesenheit eines anderen maskierenden Geräuschs (hier bandbregrenztes Rauschsignal) oder Tons, wahrzunehmen. Dies ist ähnlich zu verstehen wie der absolute Schwellwert, bei dem das maskierende Signal dem internen Rauschen entspricht. Zwicker hat die Bandbreite des maskierenden Signals so lange erhöht, bis die Mithörschwelle konstant blieb. Die Bandbreite, bei der dieser Fall eintritt, ist dann die Bandbreite der Frequenzgruppe. Unterhalb der Bandbreite, das heißt sofern die Bandbreite des maskierenden Signals innerhalb der Bandbreite der Frequenzgruppe liegt, erhöht sich die Mithörschwelle linear. Dies liefert für die obige Transformation 24 Frequenzgruppen (Bänder) für einen Frequenzbereich von 0 bis 16 kHz. Schröder kommt zu ähnlichen Ergebnissen [28].

Es ist zu bemerken, dass trotz der Einteilung in eine bestimmte Anzahl definierter Bänder die Anzahl der Bänder von dem psychoakustischen Modell abhängt und auch innerhalb des Modells mit der Abtastfrequenz stark variiert. Eine große Anzahl von Bändern (größer 24) verteilt das Rauschen schon bei niedrigen Bandbreiten auf mehrere Bänder, das heißt, eine Erhöhung der Bandbreite des maskierenden Rauschens über ein Band hinaus, müsste schon bei der jetzt geringeren Bandbreite zu einer konstanten Mithörschwelle im betrachteten Band führen. Diese Mithörschwelle müsste dann kleiner sein als die Mithörschwelle bei Modellierung mit weniger Bändern unter der Voraussetzung, dass für beide Fälle die gleiche Bandbreite des maskierenden Rauschsignals verwendet wird. Eine Superposition der Mithörschwellen eines Modells A mit vielen Bändern und der gleichen repräsentierten Bandbreite im Frequenzbereich, wie ein Modell B mit wenigen Bändern, sollte auf die gleichen Mithörschwellen im Frequenzbereich führen wie eine Superposition der Mithörschwellen des Modells B. Die Frequenzgruppen stellen keine diskreten Abschnitte der Frequenzskala dar, [43]. Die Erregung (gemeint ist die Anregung auf der Basilarmembran) ist kontinuierlich, also auch für einen Sinuston und nicht für ein breitbandiges Signal nicht

nur auf die Resonanzstelle beschränkt. Daher können sich die Frequenzgruppen praktisch an jeder Stelle bilden [43]. Es stellt sich dennoch die Frage wie sich die Anzahl der Bänder bei der praktischen Implementierung der im nachfolgenden Kapitel beschriebenen psychoakustischen Modelle auswirkt.
Weiterhin muss beachtet werden, dass das hier erwähnte Rauschen nur zur Bestimmung der Frequenzgruppen verwendet wurde und daher nicht im Zusammenhang mit den Erläuterungen der im ersten Kapitel beschriebenen Motivation zur Ausnutzung von Verdeckungseffekten steht - wie dort beschrieben soll das Sprachsignal zur Verdeckung des Rauschen genutzt werden und nicht umgekehrt.

Die Mittenfrequenzen, sowie die oberen und unteren Grenzfrequenzen finden sich in [14] und [28]. Die zu den in den Gleichungen 3.4 bis 3.7 gegebenen Transformationsvorschriften gehörenden Bandbreiten werden mathematisch durch folgende Gleichungen beschrieben.

$$\mathrm{cbw_z}(f) = 25 + 75 \cdot (1 + 1{,}4 \cdot 10^{-}6 \cdot f^2)^{0{,}69}\ , \ \text{in [Hz]} \tag{3.8}$$

$$\mathrm{cbw_s}(z) = \cosh\left(\frac{z}{7}\right) \cdot \frac{650}{7}\ , \ \text{in [Hz]} \tag{3.9}$$

Diese Gleichungen für die sogenannten kritischen Bandbreiten (Bandbreiten der Frequenzgruppen) stehen mit den Transformationsvorschriften in den Gleichungen 3.4 und 3.5 durch Integration der reziproken kritischen Bandbreitenfunktion über z in Zusammenhang:

$$\mathrm{hz2b}(z) = \int \frac{1}{\mathrm{cbw}(z)}\, dz \tag{3.10}$$

Die Funktionen sind in Abhängigkeit der Frequenz in Abbildung 3.6 aufgetragen. Die Bandbreite der Äquivalentrechteckbandbreitenskala (schwarze Linie) ist zum Vergleich auch eingezeichnet. Die Bandbreiten sind im Frequenzbereich von 1 bis 9 kHz sehr ähnlich, jedoch unterscheidet sich die Bandbreite der Approximation nach Schröder bei hohen Frequenzen von der Approximation nach Zwicker, was den Unterschieden der in Abbildung 3.5 oben gezeigten Transformationsvorschriften entspricht. Die Äquivalentrechteckbandbreite, auf der viele Filterbank basierte Modelle basieren, weist bei niedrigen Frequenzen bis zu 500 Hz deutlich geringere Bandbreiten auf als die beiden Barkskalen.

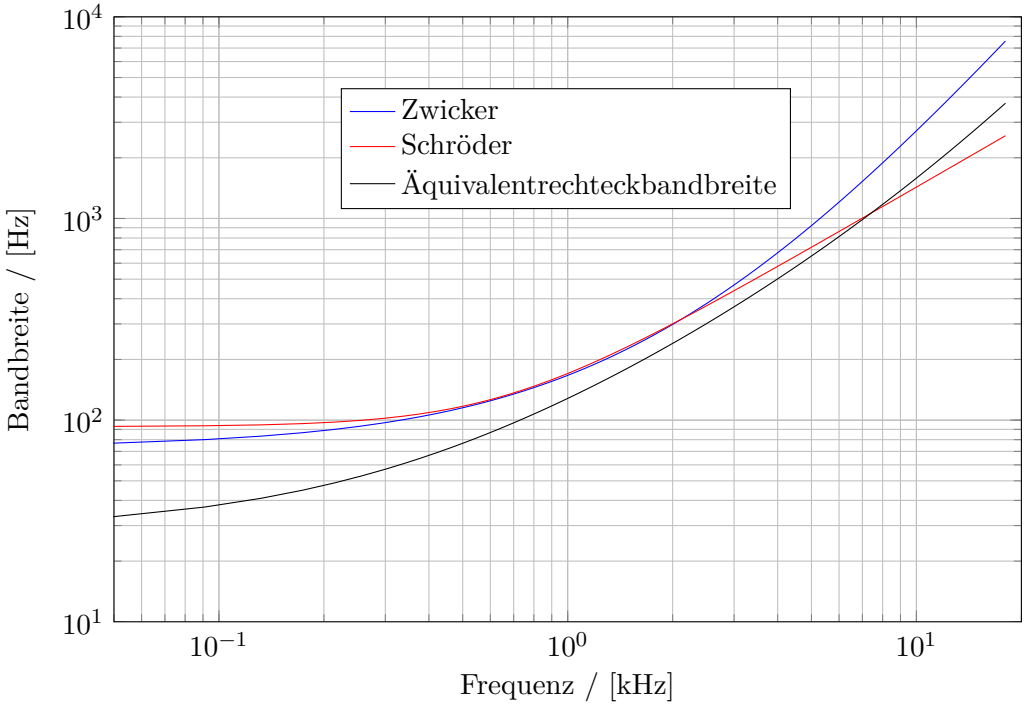


Abbildung 3.6: Bandbreite der Approximationen der kritischen Bänder des Gehörs

Als Beispiel für die Aufteilung der Frequenzachse in Frequenzgruppen zeigt Abbildung 3.7 mit 38 Bändern, deren Mittenfrequenzen als blaue Linien zu sehen sind.

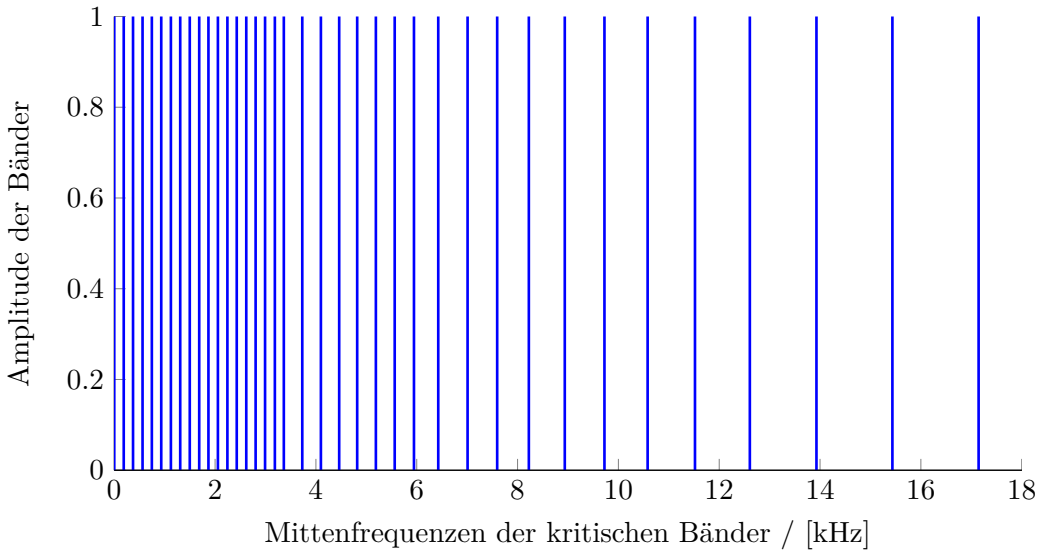


Abbildung 3.7: Mittenfrequenzen der kritischen Bänder des psychoakustischen Modells 2 des MPEG1 Layer 3 Standards [19]

Ungeachtet der Modellierung der spektralen Auflösung durch Frequenzgruppen oder

Bänder von Filterbänken steigt die Bandbreite bei hohen Frequenzen an. Für die damit einhergehende sinkende Frequenzauflösung wird das menschliche Gehör zu höheren Frequenzen hin immer weniger anspruchsvoll. Zum einen werden Phasenlagen ab etwa 2 kHz nicht mehr erkannt, zum anderen wird der Klangunterschied zwischen Sinustönen und Schmalbandrauschen zu hohen Frequenzen immer geringer. Dies kann man sich leicht vorstellen, wenn man bedenkt, dass bei niedrigen Frequenzen die Bänder sehr schmal sind, wie auch in der oberen Abbildung zu sehen ist. Die neuronale Nachverarbeitung des menschlichen Gehörs gleicht die fehlende Phasenauflösung wieder aus. Es reagiert sehr sensitiv auf zeitliche sowie auch spektrale Geräuschmuster. Das Phänomen ähnelt der Tatsache, dass man Sätze bestehend aus Wörtern mit verdrehten Buchstaben noch gut lesen kann [34].

3.2 Verdeckungseffekte

In Abschnitt 2.7 wurden die Störleistungsdichte der Sprache R_{qsqs} und des Rauschens R_{qnqn} vorgestellt. Die Mithörschwelle (auch Verdeckungsschwellwert) R_{TT} kann genutzt werden, um eine höhere Rauschleistungsdichte R_{qnqn} zuzulassen, so dass, wie in der Abbildung 2.4 gezeigt, die Störleistungsdichte der Sprache minimiert werden kann, um die Sprachverzerrung zu reduzieren, ohne dass der wahrnehmbare (d.h. hörbare) Rauschanteil am verbesserten Signal ansteigt. Ein Maß für die Verdeckung ist das Signal zu Rauschverhältnis SMR [4], welches das Verhältnis von dem maskierenden Signal (Nutzsignal) zum zu maskierenden Signal (Rauschen) beschreibt. Im Fall der Störgeräuschreduktion ist das klare Sprachsignal (oder praktisch dessen Schätzwert) das Nutzsignal, welches ähnlich wie bei den Codierungsverfahren (MP3) bei denen wahrnehmbares Quantisierungsrauschen minimiert werden soll, nun wahrnehmbares Rauschen reduziert. Das Signal- zu Maskierungsverhältnis kann man mittels des Störabstands ausdrücken. Der Störabstand ergibt sich aus der Summe des Signal zu Maskierungsverhältnisses SMR und des Rausch- zu Maskierungsverhältnisses NMR. Letzteres beschreibt den Abstand zwischen Energieniveau der Mithörschwelle und Rauschanteil.

$$SMR(f) = SNR(f) - MNR(f) \tag{3.11}$$

$$SNR(f) = SMR(f) + MNR(f) \tag{3.12}$$

Die Mithörschwelle $R_{TT}(f)$ berechnet sich dabei aus der Erregung E_x von der das Signal zu Rauschverhältnis abgezogen wird.

$$R_{TT}(f) = E_x(f) - SMR(f) \tag{3.13}$$

In der Abbildung 3.8 sind die Größen aufgetragen. Der schwarze Punkt kennzeichnet einen Sinuston auf der Tonheitsachse mit relativ hoher Energie, welcher exemplarisch die Sprache repräsentiert. Der rote Bereich stellt das schmalbandige Rauschen da. Die blaue Linie zeigt qualitativ die Mithörschwelle, welche durch den Sinuston generiert wird. In der Abbildung verdeckt der Sinuston das Rauschen vollständig. Es wird im Wesentlichen zwischen zwei Verdeckungsphänomenen unterschieden: der spektralen und der temporalen

[4] SMR: signal to mask ratio

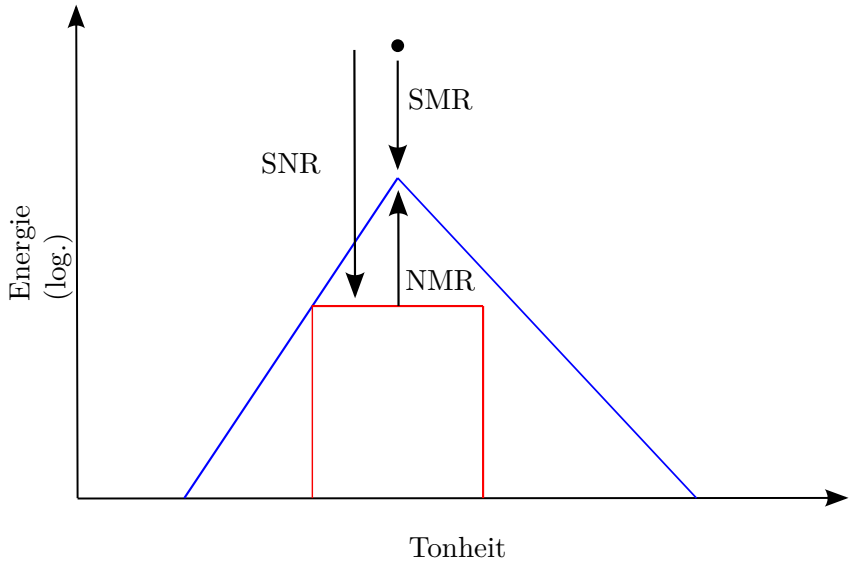


Abbildung 3.8: Signal zu Maskierungsverhältnis

Verdeckung (auch simultane und nicht simultane Verdeckung). Die spektrale Verdeckung beschreibt die Verdeckung eines Tons oder Rauschens im Spektrum, d.h. maskierendes Signal und maskiertes Signal treten simultan auf, woher auch der synonym verwendete Name „simultane Verdeckung“ rührt. Hingegen beschreibt die temporale Verdeckung Effekte, bei der Maskierer (das maskierende Signal) nicht simultan mit dem zu maskierenden Signal auftritt.

3.2.1 Spektrale Verdeckung

Es werden mehrere Typen spektraler (simultaner) Verdeckung unterschieden: Ton verdeckt Rauschen [43], Rauschen verdeckt Ton, Rauschen verdeckt Rauschen [21]. Anzumerken ist, dass die hier als spektral bezeichnete Verdeckung in der Umsetzung in geringem Maße auch Komponenten der temporalen Verdeckung enthält, da das für die Berechnung nötige Spektrum über N Abtastwerte, d.h. über eine gewisse Zeitspanne ermittelt wird. Für die Anwendung in der Störgeräuschreduktion sind nur die Verdeckungseffekte Ton zu Ton und Ton zu Rauschen interessant. Der Effekt der simultanen Maskierung und auch bis zu einem gewissen Grad der temporale Maskierung hängt stark von dem Spektrum des Verdeckers (Maskierers) als auch des verdeckten (maskierten) Signals ab [14]. Da alle Signale durch Fourierreihenzerlegung in Sinustöne zerlegt werden können, stellt der letztere Fall und auch die hier nicht relevanten Fälle nur eine Superposition vieler Ton zu Ton Maskierungen dar. Die Superposition muss dabei nicht notwendigerweise linear verlaufen.

In der Abbildung 3.8 ist anhand des qualitativen Verlaufs der Mithörschwelle zu sehen, dass die Anregung durch einen Sinuston auf der Basilarmembran wie oben erwähnt, nicht auf die Resonanzstelle begrenzt ist, sondern eine Spreizung erfährt. Die Verläufe des daraus resultierenden Erregungsmusters und der Mithörschwelle sind in weiten Teilen identisch [43], welches einen konstanten „Offset“ für die Berechnung der Mithörschwellen aus der Erregung motiviert. Der Wert für das Rausch zu Maskierungsverhältnis NMR ist

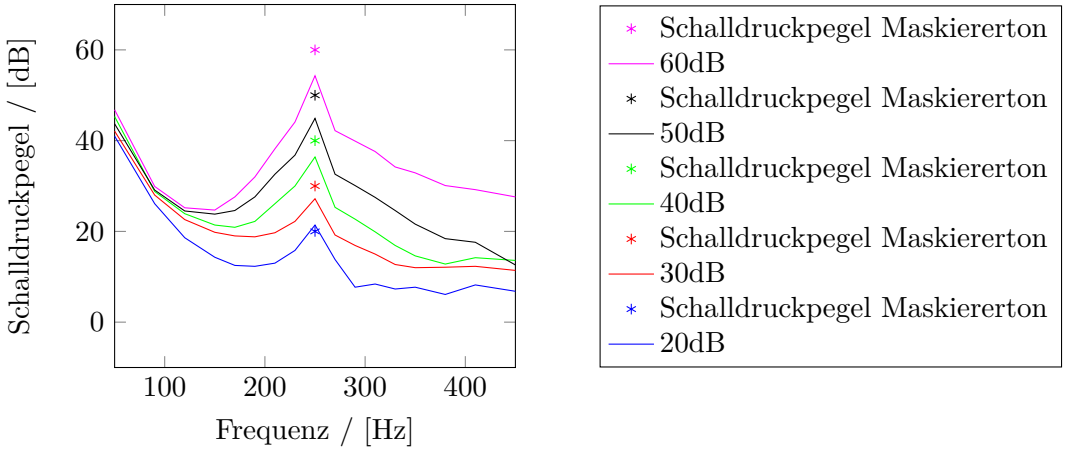
im Idealfall negativ. Für positive Werte ist das Signal zu Maskierungsverhältnis groß und die Mithörschwelle liegt unterhalb des Energieniveaus des Rauschens, so dass ein Verdeckungseffekt ausbleibt. Dies tritt bei niedrigen Störabständen auf. Die gerade bei diesen Störabständen nötige Maskierung fällt dementsprechend gering aus, woraus sich Grenzen für die Psychoakustik zur Nutzung bei der Störgeräuschreduktion ergeben.

3.2.1.1 Verdeckung eines Tons durch einen Ton

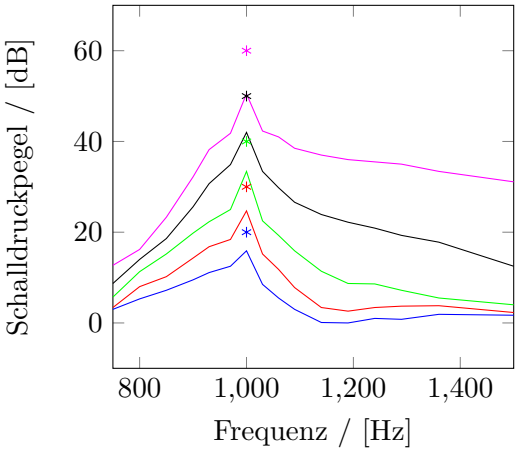
In den unteren Abbildungen 3.9a bis 3.9d sind die Mithörschwellen in Dezibel für jeweils einen Sinuston (Sternchen) und Schalldruckpegel von 20 bis 60 dB aufgetragen. Die Mithörschwellen sind jeweils in den gleichen Farben eingezeichnet, wie der diese generierende Ton. Bei niedrigen Frequenzen, exemplarisch hier durch einen Sinustons von 250 Hz in Abbildung 3.9a dargestellt, beeinflusst die oben vorgestellte absolute Mithörschwelle (Abbildung 3.4) den resultierenden Verlauf der Maskierung (links in der Abbildung 3.9a). Die durch Zwicker mittels Messungen gewonnenen Kurven zeigen, dass je höher der Schalldruckpegel des anregenden Sinustons ist, desto größer wird der Signal zu Maskierungsabstand SMR, da der Abstand zur absoluten Hörschwelle zu hohen Schalldruckpegeln immer geringer wird. Deutlich zu sehen ist auch die Asymmetrie der Mithörschwellen in Bezug auf die Frequenz. Dieses Phänomen wird als Asymmetrie der Verdeckung bezeichnet. Beim Vergleichen der Schwellen für Sinustöne niedriger Frequenz mit denen hoher Frequenz fällt auf, dass die oberen Flanken für den Maskierer niedriger Frequenz für alle Schalldruckpegel stärker abfallen als für Maskierer hoher Frequenzen (z. B.: Abbildungen 3.9a, 3.9c). Die Asymmetrie dreht bei Erhöhung der Frequenz des Maskierers sogar um. Während bei niedrigen Frequenzen die untere Flanke der Mithörschwelle steiler als die obere Flanke verläuft, ist es bei hohen Frequenzen zumindest bei höheren Schalldruckpegeln umgekehrt. Zusammenfassend kann man sagen, dass die obere Flanke der Mithörschwelle mit steigender Frequenz und auch mit steigendem Schalldruckpegel flacher verläuft, jedoch hängt die Mithörschwelle mehr von dem Schalldruckpegel des verdeckenden Signals (Maskierer) und weniger von der Mittenfrequenz des kritischen Bands (Bark Skala) ab [21]. Tonale Anregungen, deren Frequenzen nah beieinander liegen verdecken sich gegenseitig.

Dabei ist L der Schalldruckpegel in dB und f_c die Mittenfrequenz der betrachteten Frequenzgruppe. Diese Approximation ist Teil des in dieser Arbeit verwendeten psychoakustischen Modells nach [37] (Kapitel 4. In der oberen Abbildung 3.8 kann man sehen, dass die neuronale Erregung für einen Sinuston gemessen in dB ungefähr dreiecksförmig verläuft. Diese Form ist größtenteils invariant entlang der Barkskala, d.h. nahezu unabhängig von der Tonheit. Der Abfall der Mithörschwelle zu niedrigen Frequenzen hin ist größtenteils unabhängig von dem Erregungsniveau, also dem einfallenden Schallpegel. Die Steigung liegt im Bereich von 27 dB / Bark [37] [14]. In den Abbildungen 3.9c ist der Verlauf der oberen Flanke der Mithörschwelle bei zunehmender Frequenz flach bis wieder ansteigend. Während bei einem Schalldruckpegel von 20 dB der Anstieg auf den Einfluss der absoluten Hörschwelle zurückzuführen ist, liegt der Grund für das zweite Maximum bei ca. 5250 Hz der Mithörschwelle des 4000 Hz Sinustons 60 dB an der Nichtlinearität des Gehörs. Bei hohen Schalldruckpegeln treten aufgrund nichtlinearer Verzerrung wie in einem Mischer Intermodulationsprodukte aus dem Testton und Maskiererton auf wie bei dem erwähnten zweiten Maximum auf. Das gleiche Phänomen beschreibt auch Ehmer in [6]. Während Intermodulationsprodukte, welche bei Frequenzen oberhalb des Maskierers auftreten, durch

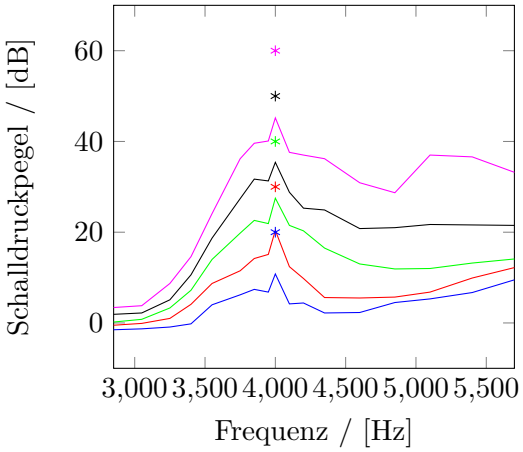
die flache Flanke der Mithörschwelle teilweise unhörbar werden, können Differenztöne mit Frequenzen unterhalb der Maskiererfrequenz durch die steile untere Flanke der Mithörschwelle nicht ausreichend verdeckt werden [14]. Weiterhin können auch Produkte entstehen, die die Mithörschwelle absenken [12]. Dies erklärt auch den Abfall der Mithörschwelle in Abbildung 3.9c) bei ca. 4800 Hz mit anschließendem Anstieg. In Abbildung 3.9e sind exemplarisch die Maskierungsschwellen für die Frequenzen 0,45, 1 und 4 kHz nicht über der Frequenz sondern der Tonheit aufgetragen. Wie in Abbildung 3.8 angedeutet, ist der Verlauf ausgenommen der Intermodulationsmaxima nahezu dreieckförmig und näherungsweise invariant entlang der Tonheitsskala. Die Betrachtung der Verläufe im Bereich der kritischen Bänder (Tonheit) führt also zu einer Linearisierung.



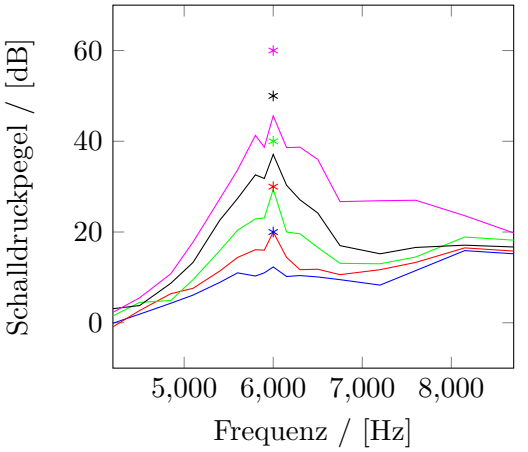
(a) Sinuston mit einer Frequenz von 250 Hz



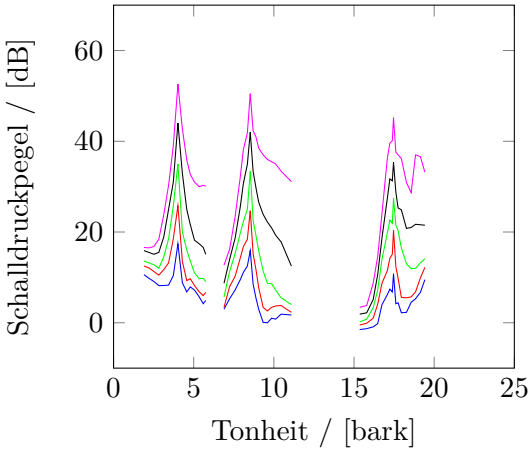
(b) Sinuston mit Frequenz von 1000 Hz



(c) Sinuston mit Frequenz von 4000 Hz



(d) Sinuston mit Frequenz von 6000 Hz



(e) Sinustöne 0,45, 1, 4 kHz in Bark

Abbildung 3.9: Maskierungschwellwerte in Abhängigkeit des Schalldruckpegels für verschiedene Sinustöne, Daten aus [40]

Die resultierende sogenannte „globale“ Mithörschwelle R_{TT} berechnet sich aus der Superposition der durch die verschiedenen im Sprachsignal vorkommenden Sinustöne erzeugten Mithörschwellen. Da diese sich nicht lokal auf ein Band beschränkt, sondern über alle Bänder erstreckt, wird sie als „globale“ bezeichnet. Bei der nichtlinearen Addition der Erregungsmuster liegt die globale Mithörschwelle bei mehreren Maskierern in der Regel höher als die Summe der einzelnen Mithörschwellen [38]. Der Wert beschreibt das gerade hörbare Niveau des Rauschens (engl: just noticable noise, kurz JND) in Abhängigkeit der Frequenz oder der Tonheit und wird auch als Modifikation des Frequenzverlaufs der absolute Hörschwellen des Schalldruckpegels gesehen [12], welches auch auf der Idee der Isophone, d.h. Kurven gleich empfundener Lautstärke in Abschnitt 3.3.1 führt. Die Superposition wird in verschiedenen Ansätzen zur Modellierung als linear und nichtlinear beschrieben. Die oben erwähnten nicht linearen Phänomene sowie die Nichtlinearitäten des menschlichen Gehörs [14] und [24] motivieren eine nicht lineare Superpostion. Lutfi [24] hat ein exponentielles Superpositionsgesetz für die Erregungen aufgestellt, aus denen sich zu diesen vorwiegend parallel verlaufende Mithörschwellen berechnen. Die globale Erregung E_{xg} ergibt sich sich aus den einzelnen Erregungen E_x der Bänder b

$$E_{xg} = \left(\sum_b E_x^{\alpha} \right)^{\frac{1}{\alpha}} \tag{3.14}$$

Lutfi [24] gibt für den Exponenten α 0.4 an.

3.2.1.2 Ton maskiert Rauschen

Im Allgemeinen wird Rauschen durch einen Ton weniger verdeckt als ein Ton durch einen Ton [14]. Allerdings gilt, solange sich das Rauschen mit seinem kompletten Spektrum unterhalb der Mithörschwelle befindet, ist es unhörbar. Oft liegt der Fall vor, dass Teile des Rauschens, welches eine gewisse Bandbreite aufweist, oberhalb der Mithörschwelle liegen und daher wahrnehmbar sind. Der Fall kann, sofern die Schalldruckpegel niedrig und daher nichtlineare Verzerrungen unwahrscheinlich sind, als Superposition vieler Ton zu Ton Betrachtungen gesehen werden. Es werden im Allgemeinen andere Signal zu Maskierungsverhältnisse SMR von den Erregungsenergien zur Berechnung der Maskierungsschwellen verwendedet als im Ton zu Ton Fall. Als oberen und unteren Grenzwert existieren der sogenannte „Rausch zu Ton Maskierungsabstand“ (engl. „Noise masking Tone“ - kurz: NMT) und der „Ton zu Rauschen Maskierungsabstand“ (engl. „Tone masking noise“ - kurz: TMN). Für diese auch als „Offset“ bezeichnete Größen existieren zahlreiche Werte in der Literatur. Für letzteren bis zu 28 dB [19]. Beide Werte setzen eine Charakterisierung des Maskierers und des maskierten Signals (Rauschen) voraus, welches dann zu einer Spektralanalyse wie das spektrale Flachheitsmaß nach [34] oder den Tonalitätsindex aus [19] führt. Bei letzterem wird allerdings nur der Maskierer analysiert und lässt eigentlich keine Klassifizierung in eine Rauschen maskiert Ton oder Ton maskiert Rauschen Situation zu. Bei der Maskierung von Rauschen durch Töne ist die Größe TMN abhängig von der Tonheit z und lässt sich mit folgender Gleichung aus [38] ausdrücken:

$$TMN = 15.5 + z\,, \quad \text{in [dB]} \tag{3.15}$$

Dabei ist die Einheit von z Bark und der resultierende Wert in Dezibel. Andere Werte für manche Bänder teils ähnliche Wert lassen sich in [5], [17] und [28] finden. Der minimale Signal- zu Verdeckungsabstand TMN, beträgt 21 bis 28 dB.

3.2.1.3 Rauschen maskiert Ton

Verdeckung durch Rauschen (breitbandige Signale) ist aufgrund der höheren Gesamtenergie (unter Vorausetzung des gleichen Schalldruckpegels) stärker als durch tonale Signale [21]. Ist die Frequenz eines Tones, der durch Rauschen verdeckt wird, nahe der Mittenfrequenz des Rauschens und 5 dB unterhalb des Pegels des Rauschsignals, so ist der Ton unhörbar, [37]. Daraus ergibt sich für den im Abschnitt erwähnte Größe NMT ein Wert in dieser Größenordnung [19], [32]. In diesem Fall ist das Ausmaß der Maskierung nahezu unabhängig von der Frequenz des Rauschens.
Allgemein gilt jedoch für die spektrale Verdeckung, dass das Ausmaß der Verdeckung im Wesentlichen von der Struktur des Maskierers und dem relativen Frequenzabstand des Maskierers und des maskierten Signals bestimmt wird [37].

Die maximale Verdeckung liegt bei der Verdeckung von Tönen durch Rauschen unabhängig von der Mittenfrequenz des Rauschens ca. 5 dB unter dem Pegel des Maskierers.

Zusammendfassend kann man sagen, dass eine tonale Verdeckung zu einer asymmetrischen Maskierungskurve auf der Frequenzachse, und eine breitbandige Maskierung zu einer zum Maskierungsfreuquenzbereich symmetrischen Maskierungskurve führt [6].

3.2.2 temporale Verdeckung

Die temporale Verdeckung lässt sich in die Vorverdeckung und in die Nachverdeckung einteilen. Die Vorverdeckung beschreibt die nachträgliche Verdeckung eines schon in das Ohr eingetretenen Schalls durch einen danach auftretenden Schall (den Maskierer). Die Nachverdeckung bezeichnet das Phänomen der nach Auftreten des Maskierers noch vorhandenen Restmithörschwelle, welche mit der Zeit abklingt und währenddessen Signale verdeckt. In der unteren Abbildung ist die Mithörschwelle im zeitlichen Verlauf vor (negative Zeitachsenwerte), während (erste Skala von 0 bis 200 ms) und nach Auftreten des Maskierers (auf der Skala mit zweiter Null beginnend) zu sehen. Während des Erklingen des Maskierers folgt die Mithörschwelle den im vorigen Abschnitt erwähnten Phänomenen. In der Abbildung ist zu sehen, dass die Verdeckung bzgl. der Zeit asymmetrisch ist. Die Vorverdeckung ist mit maximal 20 ms Dauer (typischer Wahrnehmung: 5 ms für hohe Amplituden des Verdeckers [14], [26], [29] und weitaus geringer ausgeprägt als die Nachverdeckung mit einer Dauer von bis zu 200 ms [5], [21], [26].

3.2.2.1 Vorverdeckung

Die Vorverdeckung ist darauf zurückzuführen, dass das Gehirn Gehörtes zwischenspeichert (bis zu 20 ms) und innerhalb dieser Zeitspanne eine Priorisierung und Datenreduktion bei der Verarbeitung verschiedener Schallereignisse durchführt. Die Verarbeitung von Audiosignalen hoher Lautstärke erfolgt im Gehirn schneller als mit geringer Lautstärke . Daher können schwächere Signale weniger hörbar werden, [37], [38]. Zwicker beschreibt in [14] einen Anstieg der Mithörschwelle auf das Niveau der simultanen Verdeckung in einer der

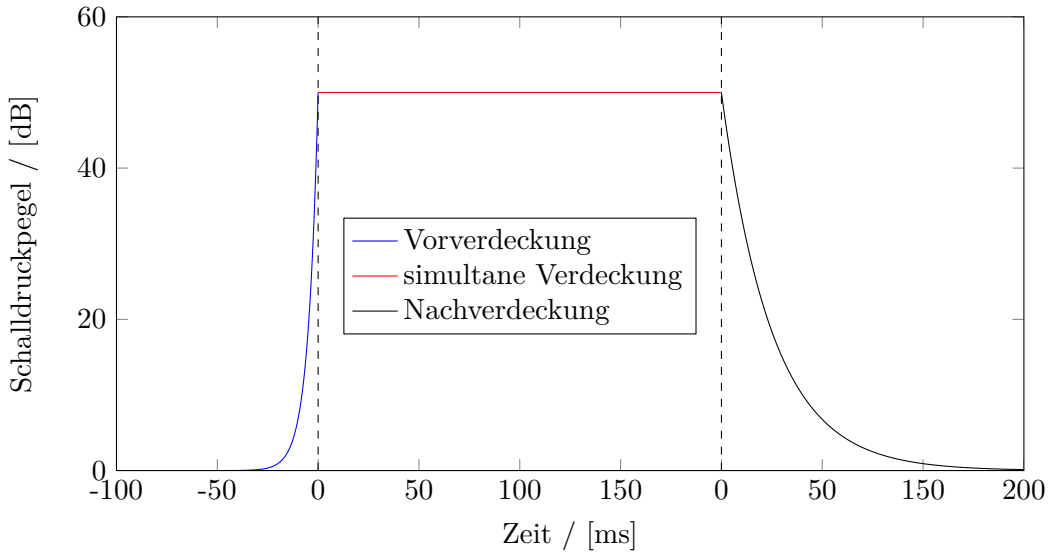


Abbildung 3.10: temporale Verdeckung, Daten aus [14]

oben erwähnten Zeitspanne, welche unabhängig von dem Pegel des Maskiertons ist. Vorverdeckung tritt bei geübten Hörern kaum bis gar nicht auf [4], [26] und auch in [14] wird das Phänomen gegenüber der deutlich länger wirkenden Nachverdeckung als nahezu vernachlässigbar angesehen. Vorverdeckung ist bei niedrigen Frequenzen stärker ausgeprägt als bei hohen Frequenzen, vorrausgesetzt der Schalldruckpegel ist gleich. Das Ausmaß der Verdeckung wird durch die Frequenzdifferenz zwischen Maskierer und Signal beeinflusst [27]. Die Mithörschwelle bei der Vorwärtsverdeckung ist höher, je näher der betrachtete Zeitpunkt am Anfang des Maskierungssignals liegt.

3.2.2.2 Nachverdeckung

Der Effekt der Nachverdeckung wird durch verschiedene Effekte wie zeitliche Überlappung der Antworten der Basilarmembran oder auch kurzzeitige Ermüdung (Fatigue) des höheren Nervenssystems und Nachwirken des neuronalen Aktivität bestimmt. Entsprechend der Erholungszeit bei den Erregungsmustern werden leise Signale kurz nach lauten Signalen durch diese unhörbar. Die Höhe des Verdeckungsschwellwerts steigt mit der Dauer des Maskierers, der Amplitude und Verringerung der zeitlichen Verzögerung zwischen Maskierer und maskiertem Signal [14]. Der Verdeckungseffekt nimmt bei zunehmenden zeitlichen Abstand zwischen Maskierer und maskiertem Signal ab. Die Erholungsrate der Nachverdeckung ist größer für höhere Maskieramplituden, daher fällt die Maskieramplitude ungeachtet ihrem Anfangsniveau nach 100 bis 200 ms auf Null ab [27]. Bei simultaner Verdeckung mit breitbandigen Maskierern ist das Signal zu Maskierungsverhältnis (SMR) konstant. Hingegen gilt dies nicht für nicht-simultane Verdeckung. Eine Erhöhung der Amplitude des Maskierers führt nicht notwendigerweise zur gleichen Erhöhung des Maskierungsschwellwerts. Die Maskierungsschwellwerte sind laut Untersuchungen von Moore proportional zum Logarithmus der Verzögerungszeit in [27] dargestellt.

3.3 Empfindungsgrößen

In Untersuchungen [43] hat sich gezeigt, dass der Schalldruckpegel als Messgröße der menschlichen Empfindung von Lautstärke nicht gerecht wird. Obwohl zwei Schalle unterschiedlicher Frequenz oder Struktur (bei komplexeren Geräuschen) den gleichen Effektivwert des Schalldruckpegels aufweisen, werden sie als unterschiedlich laut wahrgenommen. Daher wurden die Größen Lautstärke und Lautheit eingeführt. Während die oben genannte Tonheit die Wahrnehmung der Frequenz beschreibt, stellen diese Größen dies für den Schalldruckpegel dar.

3.3.1 Lautstärke

Die Idee der Lautstärke besteht darin, wie in der internationalen Konvention festgelegt, Schall mit einem Referenzschall, dem sogenannten „Standardschall“ zu vergleichen. Das in der Einheit phon gemessene Maß ist für ein 1 kHz Signal und 40 dB Schalldruck auf ein phon festgelegt. Daher stellt es wegen der Verknüpfung zum Schalldruckpegel ein halb objektives und wegen der subjektiven Bewertung eines Schalls durch Vergleich ein halb subjektives Maß dar. In der unteren Abbildung 3.11 sind die Kurven konstanter gehörrichtiger Lautstärke nach [18] dargestellt. Für die Frequenz von 1 kHz entspricht wie mit der senkrechten roten Linie gekennzeichnet der phon Wert dem Schalldruckpegel in Dezibel. Beim Vergleich mit der absoluten Hörschwelle in der Abbildung 3.4 oben fällt auf, dass die Kontouren relativ parallel zu diesem verlaufen und zu hohen Schalldruckpegeln immer flacher verlaufen. Die Lautstärke eines Schallereignisses mit einer Frequenz von 20 Hz wird um 80 dB niedriger empfunden als die eines Schalls mit gleichem Schalldruckpegel und einer Frequenz der größten Hörempfindlichkeit bei ca. 3,4 kHz.

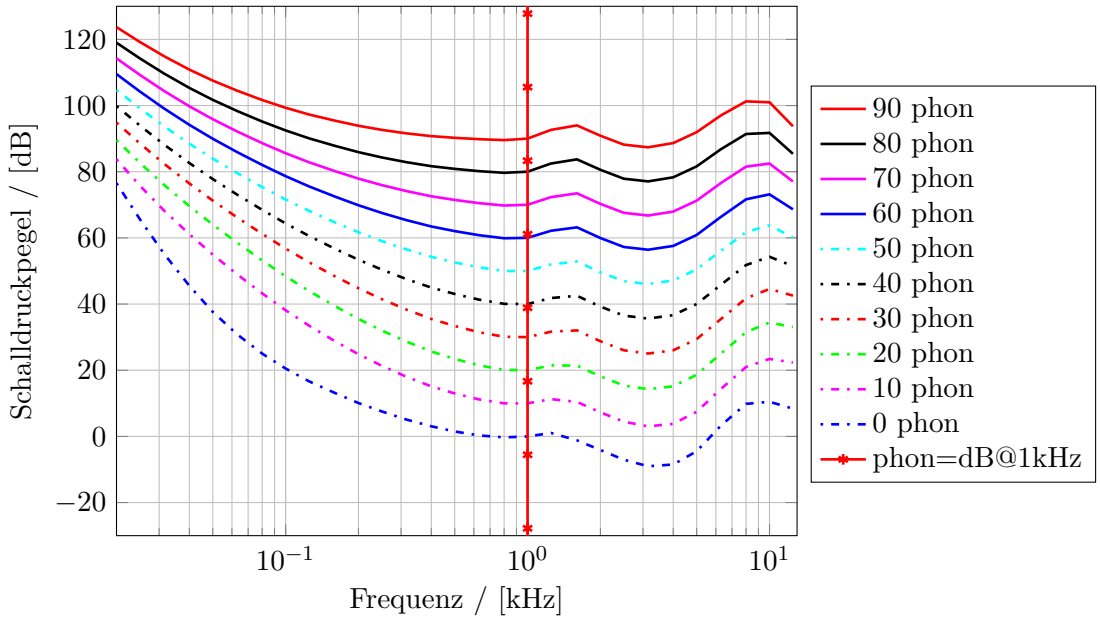


Abbildung 3.11: Isophone (Kurven konstanter gehörrichtiger Lautstärke) nach ISO 226:2003 Standard [18]

3.3.2 Lautheit

Während die Lautstärke die Wahrnehmung von Schallereignissen auf den 1 kHz Ton referenziert, beruht der Grundgedanke der Einführung der Größe „Lautheit“ darauf, die wahrgenommene Lautstärke von verschiedenen Ereignissen relativ miteinander zu vergleichen, d.h. wie viel mal lauter oder leiser ein Ereignis zu einem anderen ist [43]. Als grobe Näherung gilt, ein 10 dB höherer Schalldruckpegel wird als doppelt so laut, ein 10 dB niedriger Schalldruckpegel als halb so laut wahrgenommen [11]. Das Gehör wird mit wachsendem Schalldruck immer empfindlicher gegenüber Amplitudenänderungen von Sinustönen [43]. Bei einem niedrigen Schalldruckpegel von 20 dB liegt der eben wahrnehmbare Modulationsgrad (also eine wahrnehmbare Änderung der Lautheit) bei einem Wert von etwa 10 % des Schalldruckpegels, bei einem Pegel von 100 dB erreicht er etwa den Wert von 1 % [43]. Die Dauer des Klangs bestimmt die wahrgenommene Lautstärke, je länger die Dauer des Klangs, desto lauter wird er wahrgenommen [11], [30]. In Experimenten [30] wurde gezeigt, dass ein Rauschimpuls mit konstant definierter Leistung als lauter wahrgenommen wird, sobald dessen Bandbreite die des angeregten kritischen Bands überschreitet [11]. Die empfundene Lautstärke hängt also nicht nur vom Schalldruckpegel, sondern auch von der Dauer und den temporalen und spektralen Eigenschaften des Audiosignals ab [14]. Beispielsweise werden breitbandige Schalle lauter empfunden als schmalbandige, auch wenn in beiden Fällen der effektive Schalldruckpegel gleich ist. Für Klänge bis zu einer Dauer bis zu 200 ms ist die wahrgenommene Lautstärke von der Dauer abhängig und ansteigend. Ähnlich dem Effekt der Nachverdeckung stellt dies eine zeitliche Integration der Anregung dar. Darüber hinaus ist die Frequenzabhängigkeit der Lautheit selbst abhängig von der Amplitude der Anregung [11]. Aus den oben genannten Eigenschaften wird ersichtlich, dass die Lautheit ein komplexes subjektives Maß darstellt. Ist die Wahrnehmung eines Geräuschs doppelt so laut, ist auch der sone Wert doppelt so hoch. Als Näherung für einen Zusammenhang der Lautheit N mit dem Schalldruckpegel L_p wird in Relation 3.16 angeben:

$$N \propto {L_p}^{0,6} \tag{3.16}$$

Die Empfindungsgrößen Lautheit N und Lautstärke L_N sind über die Gleichung 3.17 verknüpft.

$$N = 2^{\left(\frac{L_N - 40}{10}\right)} \tag{3.17}$$

Die Abhängigkeit ist in Abbildung 4.6b zu aufgetragen.

Es gibt mehrere Ansätze die Lautheit als empfundenen Lautstärke zu berechnen. In [20] wurden mehrere Verfahren verglichen. Die besten Ergebnisse liefern die Verfahren von Moore und Glasberg [2] und das von Zwicker [42]. Letzteres gilt dabei nach [20] als am weitesten ausgereift und soll daher im folgenden Unterabschnitt erwähnt werden.

3.3.3 Lautheitsberechnung auf Basis des Erregungsmusters

Bei der Empfindung von Lautstärke wird vom Gehör der gesamte Tonheitsbereich (die komplette Barkskala) ausgewertet [43]. Im Allgemeinen stellt ein Audiosignal eine Superposition vieler Sinustöne dar, welche zur Anregung der Basilarmembran an verschiedenen

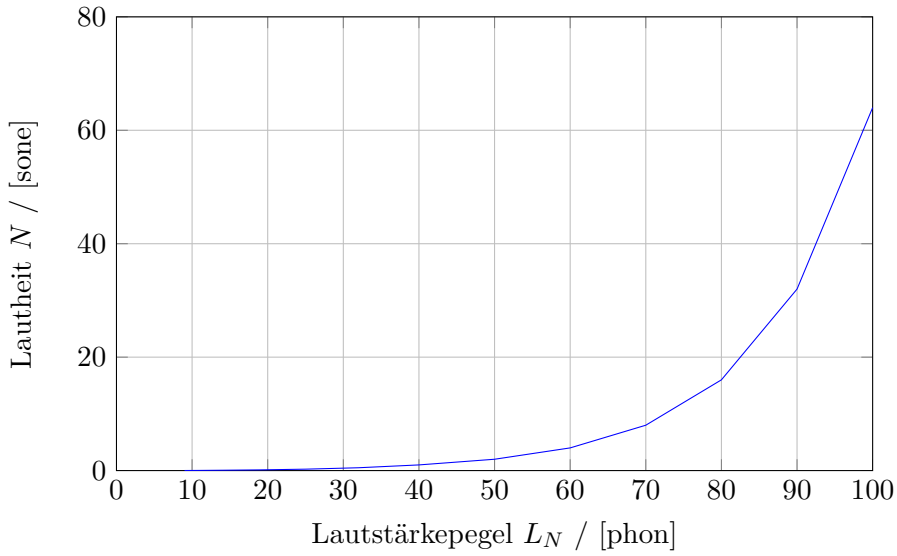


Abbildung 3.12: Lautheit vs. Lautstärke [43]

Stellen, bzw. in verschiedenen Bändern der Tonheitsskala führt. Selbst für den Fall, dass nur ein einzelner Sinuston das Innenohr erreicht, führt dies nicht nur zur Erregung der Basilarmembran an der Resonanzstelle, sondern auch in benachbarten Bereichen wie in den Abbildungen 3.9 zu sehen ist. Diese benachbarten Bereiche und auch die Resonanzstelle selbst können in der diskreten Unterteilung der Frequenzskala durch die Tonheitsskala auf mehrere Bänder fallen, d.h. die Erregung erstreckt sich somit über mehrere Bänder. Die Lautheit setzt sich demnach aus Teillautheiten (auch spezifische Lautheiten) der Bänder zusammen. Die spezifische Lautheit N' gibt die Lautheit für das betrachtete gehörrichtige Band b an. In der Gleichung 3.18 findet das Interne Rauschen IR Berücksichtigung. Dieses stellt die untere Schwelle der Erregung auf der Basilarmembran dar, welche überschritten werden muss, damit etwas „gehört“ wird. Sie beschreibt damit den Beitrag des Innenohres zur absoluten Hörschwelle aus Gleichung 3.3 Die spezifische Lautheit N' wird in der Einheit sone angegeben. Da ein sone eines 1 kHz Sinustons per Definition einem Schalldruckpegel von 40 dB entspricht, wird die Referenzgröße E_0 auf diesen Wert, bzw. den entlogarithmierten Wert 10^4 festgelegt.

$$N'(b) = 0.068 \left(\frac{1}{s(b)} \cdot \frac{IR(b)}{E_0} \right)^{0.23} \cdot \left[\left(0.5 + 0.5 \cdot \frac{E_x(b)}{E_0} \right)^{0.23} - 1 \right] , \text{ in [sone]} \tag{3.18}$$

Dabei beschreibt $s(b)$ den frequenzabhängigen Schwellenfaktor (auch Schwellenfakor der Frequenzgruppe b). Dieser gibt das Verhältnis von wahrgenommener Lautheit eines Testtons an, welcher den absoluten Schwellwert repräsentiert, also gerade eben wahrnehmbar ist und der Lautheit, welche aus dem Internen Rauschen resultiert [43]. Der Wert wird durch Einsetzen der Mittenfrequenz $f = f_c$ des Bandes b in die untere Gleichung berechnet. Für tiefe Frequenzen ergibt sich ein Wert von 0,65 , für hohe Frequenzen ein Wert von 0,25.

$$s(b) = 10^{0{,}1 \cdot \left(-2 - 2{,}05 \cdot \arctan\left(\frac{f}{4\,\text{Hz}} \right) - 0{,}75 \cdot \arctan\left(\left(\frac{f}{1{,}6\,\text{Hz}} \right)^2 \right) \right)} \tag{3.19}$$

Die resultierende wahrgenommene Lautstärke als Superposition der spezifischen Lautheiten der einzelnen Bänder wird als die „gesamte Lautheit“ N oder vereinfacht auch als „Lautheit“ bezeichnet. Diese ergibt sich aus der Integration über die gesamte Tonheitsskala bzw. der Summation über die Anzahl der Bänder Z:

$$N = \frac{24}{Z} \cdot \sum_{b=0}^{Z-1} N' \tag{3.20}$$

Der Faktor 24 in der oberen Gleichung resultiert aus der Einteilung des hörbaren Spektrums in 24 Bänder. Die Gleichung wurde entsprechend entwickelt, eignet sich aber auch für Modelle mit mehr oder weniger Bändern. Dieser Sachverhalt ist in der Variablen Z repräsentiert.

3.4 Zusammenfassung Psychoakustik

Die Eigenschaften des menschlichen Gehörs lassen sich wie folgt zusammenfassen und sollten bei der Wahl von psychoakustischen Modellen nach Möglichkeit weitestgehend berücksichtigt werden:

- Außen- und Mittelohr nehmen eine spektrale Gewichtung entsprechend der Inversen der Isophone vor.
- Das Innenohr (Basilarmembran und Haarzellen) kann als Spektrumanalysator angesehen werden [22]. Daraus ergeben sich folgende Eigenschaften
  - Frequenzauflösung entsprechend der Tonheit oder Äquivalentrechteckbandbreite
  - spektrale Verdeckungseffekte durch Spreizung der Erregung auf der Basilarmembran
  - das Erregungsmuster ergibt sich aus der Superposition einzelner Erregungen
  - Temporale Verdeckung (Integration) die zu Vorverdeckung und Nachverdeckung im Zeitbereich führen, wobei letztere deutlich bedeutsamer ist und erstere ggf. vernachlässigt werden kann.
- Für die Verdeckung gilt Folgendes
  - Je höher die Amplitude des Tones, desto stärker die Verdeckung.
  - die Mithörschwellen fallen in höheren Bändern zu hohen Frequenzen hin und auch mit höheren Schalldruckpegeln immer weiter ab.
  - der Abfall der Mithörschwellen ist zu niedrigen Frequenzen hin nahezu konstant und unabhängig von der Frequenz des Maskierers.
  - An den Rändern des Frequenzbereichs der größten Empfindlichkeit steigt sowohl zu hohen wie auch niedrigen Frequenzen die absolute Hörschwelle an und beeinflusst damit mehr und mehr die resultierende globale Mithörschwelle. Signalanteile unterhalb dieser tragen nicht zur Verdeckung bei.

  - Die Höhe der Verdeckung ist im Wesentlichen von Struktur , Signalpegel, und dem relativen Frequenzabstand des Maskierers und des maskierten Signals bestimmt.
  - Signale mit hohe Bandbreite verdecken stärker als Signale niedriger Bandbreite
  - Je länger die Dauer des Maskierertons, desto höher das Ausmaß der Nachverdeckung.

- Die empfundene Lautstärke wird durch die Lautheit beschrieben, welche unter Verwendung von psychoakustischen Modellen die Erregung in den Bändern und damit die spezifische Lautheit berechnen. Der resultierende Lautstärkeeindruck entsteht durch Integration der spezifischen Bänder über das gesamte Hörspektrum.

Kapitel 4

Psychoakustische Modelle

Im vorhergehenden Kapitel wurden aus der Psychoakustik resultierende Anforderungen für die psychoakustischen Modelle aufgestellt. Die in diesem Kapitel betrachteten Modelle sollen möglichst detailgetreu die Eigenschaften des menschlichen Gehörs mathematisch abbilden. Daher werden die Modelle am Ende des Kapitels in Bezug auf die Anforderungen bewertet.

Die Modelle dienen je nach verwendeter psychoakustische Filterregel H_{psycho} zur Berechnung der Erregungsmuster E_x, der daraus ableitbaren Verdeckungsschwellwerte [1] R_{TT} oder der Lautheit N . Letztere wurde im Kapitel 2 und Abbildung 2.5 als Teil der zweistufigen Störgeräuschreduktion vorgestellt.

Insgesamt werden vier Modelle verwendet:

1. Das erste Modell ist dem MPEG1 Layer 3 Standard [19] entnommen, im Standard wird es als „psychoakustisches Modell 2“ bezeichnet. Gustafsson [13] hat unter Einsatz von Teilen dieses Modells eine zweistufige Störgeräuschreduktion erstellt. Das Modell und die Störgeräuschreduktion mit dem schon in 2 vorgestellten Filtergewicht nach Gleichung 2.28 dient als Referenz für die in diesem Kapitel vorgestellten Modelle und die im folgenden Kapitel behandelten Filterregeln. Das Modell wird im folgenden mit MP3G abgekürzt.

2. Das zweite Modell basiert auf dem erweiterten und angepassten Modell MP3G nach Gustafsson und stellt zusätzliche Erweiterungen hier zu bereit. Die Variante mit temporaler Verdeckung und absoluter Hörschwelle wird mit MP3GADV abgekürzt.

3. Das dritte psychoakustische Modell (kurz: PEAQFFT) basiert aus dem FFT Modell des PEAQ (perceptual evaluation of speech quality) [37] Standards zur Bewertung der Audioqualität.

4. Das Filterbank basiert Modell (PEAQFB) des gleichen Standards wird als viertes Modell kurz zuletzt behandelt.

[1] genau genommen handelt es sich um Schwellwertenergien oder Leistungsdichtespektren der Verdeckungsschwellwerte

Alle hier vorgestellten Modelle bestehen aus Teilen der jeweiligen Standards und eigenen Änderungen und Erweiterungen. Letztere sind nötig, da die in den Standards beschriebenen Implementierungen nicht zur Nutzung in einer Störgeräuschreduktion dienen. Soweit nicht explizit darauf hingewiesen wird, sind die Gleichungen der Modelle entsprechend der Standards umgesetzt.

Allen Modellen gemeinsam sind folgende Verfahrensschritte, wenn auch die Reihenfolge bei dem Filterbank-Modell (Abschnitt 4.4) abweicht.

- Das aus der konventionellen Störgeräuschreduktion der ersten Stufe gewonnene (geschätzte) Sprachsignal wird in energetische Darstellung überführt.
- Die Energie wird in den Bereich kritischer Bänder transformiert.
- Die Energien werden mit einer Spreizfunktion $S(i,j)$ gefaltet. Der Index i repräsentiert das Band des Maskierers, j den Index des maskierten Bandes.
- Die gespreizten Energien werden innerhalb jedes Bandes überlagert und anschließend normiert. Diese stellen die Erregungsmuster dar.
- Ein ggf. parallel berechneter oder konstanter logarithmischer Signal zu Maskierungsabstand (SMR) wird von den gespreizten Energien subtrahiert, woraus die Verdeckungsschwellwerte R_{TT} resultieren.

4.1 Modell 1: Referenzmodell MP3G

4.1.1 Modellbeschreibung

In der unteren Abbildung 4.1 ist das Blockdiagramm des psychoakustischen Modells 2 nach Gustafsson dargestellt. Die Notation der Größen ist zum einfachen Nachvollziehen an den Standard angelehnt. Das Spektrum $\hat{S}_1(\Omega)$ des geschätzten Sprachsignals wird entsprechend der oben genannten Schritte in die Erregungsmuster umgerechnet. Die Transformation vom Frequenzbereich in den Bereich der kritischen Bänder erfolgt mit der zu der im vorhergehenden Kapitel weitestgehend übereinstimmenden Transformationsvorschrift nach Zwicker (Gleichung 3.4). Die 61 Bänder [2] ergeben sich aus einer einfachen Frequenzgruppierung und sind nicht überlappend. Parallel zur Berechnung der Erregungsmuster E_{xg} erfolgt auf Basis des Leistungsdichtespektrums LDS und der Phase des Spektrums die Bestimmung des sogenannten Tonalitätsindexes $t_i(b)$. Dieser dient dazu, das geschätzte Sprachsignal bzgl. seiner spektralen Struktur zu klassifizieren. Der Tonalitätsindex kann Werte zwischen 0 für ein breitbandiges Signal und 1 für ein tonales Signal annehmen. Der “Tonality Index“ gibt ein „likelihood“ Maß an, welches bestimmt, ob die betrachtete Komponente eher Rausch- oder tonales Verhalten aufweist. Der Index beschreibt eine Funktion der Unvorhersehbarkeit (Definition nach [5]) der spektralen Komponenten im betrachteten Rahmen. Im Allgemeinen sind tonale Komponenten besser vorhersehbar als breitbandige Signale [15]. Sie weisen daher einen höheren Wert beim Tonalitätsindex auf.

[2] für eine Abtastfrequenz von 48 kHz

Entsprechend dieses Indexes wird das Signal zu Maskierungsverhältnis SMR (im Standard als „Offset“ bezeichnet) über eine Gewichtung der Werte für den „Rauschen maskiert Ton“ (NMT) oder den „Ton maskiert Rauschen“ Fall (TMN) bestimmt. Das Signal zu Maskierungsverhältnis nimmt also Werte zwischen 5 und 29 dB an, welche vom Erregungsmuster subtrahiert werden. Der resultierende geschätzte globale Verdeckungsschwellwert (Mithörschwelle) R_{TT} wird in den Frequenzbereich zurück transformiert.

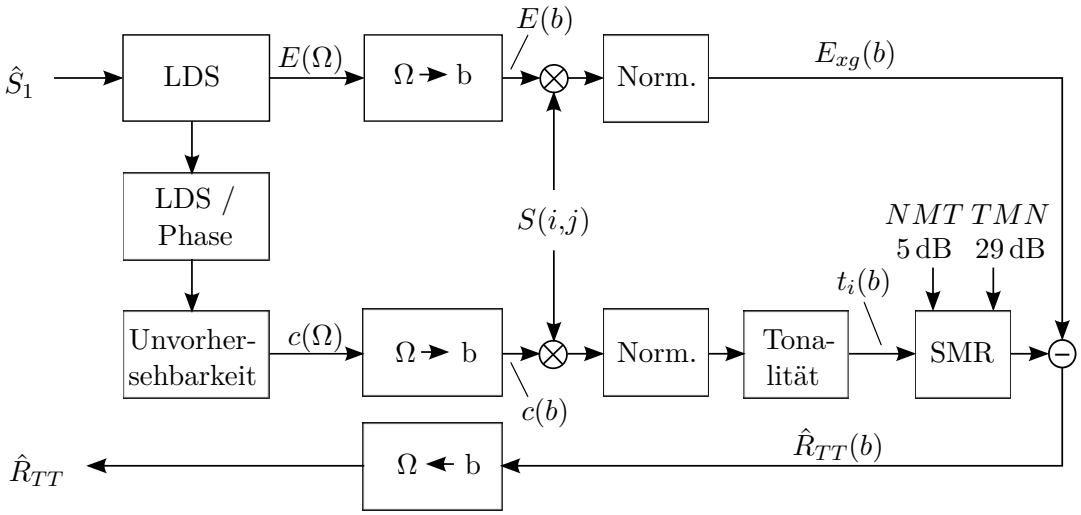


Abbildung 4.1: Implementation des psychoakustischen Modells 2 des MPEG1 Layer 3 Standards nach Gustafsson [13]

4.1.2 Anpassungen

Da einige Punkte des Standards in der Beschreibung von Gustafsson zur Implementierung [13] keine Erwähnung finden, die Berechnung des Spektrums $\hat{S}_1(\Omega)$ des geschätzten Sprachsignals nicht dem Standard und auch nicht auf im verwendeten MATLAB Framework genutzte Abtastraten anwendbar ist, werden folgende Anpassungen vorgenommen.

4.1.2.1 Transformation des Zeitsignals in den Frequenzbereich

Anzumerken ist, dass das geschätzte Spektrum durch eine DFT mit einer Länge von 1024 Abtastwerten und einer Rahmenlänge von 512 sowie einer Überlappung von 75 % gewonnen wird. Diese Werte stimmen nicht mit der ursprünglichen Implementation durch Gustafsson überein – vermutlich wurden aufgrund der von ihm verwendeten niedrigeren Abtastfrequenz des Eingangssignals die Werte angepasst.

4.1.2.2 Pre-Echo Control

Betrachtet wird hier das unter Punkt m) im psychoakustischen Modell 2 des Standards [19] beschriebene sogenannte „pre-echo control“. Aus der Darstellung von Gustafsson [13] geht nicht hervor, ob dieser Teil des Standards implementiert wurde.

„Pre-echo“ tritt auf, wenn ein Ereignis mit plötzlich hohem Signalpegel (steiler Flanke wie z. B. beim Paukenschlag) am Ende des Transformationsblocks (Rahmen) auf einen Zeitbereich mit niedrigem Signalpegel (z.B. Ruhe) folgt, der Signalverlauf also transient ist. Die Zeit-Frequenz Unsicherheit besagt, dass die inverse Transformation die Störung (durch zu geringe Zeitauflösung des transienten Ausschlags) zeitlich gleich auf den rekonstruierten Rahmen des Signals im Zeitbereich verteilt. Dabei hört man ein vorzeitiges Echo vor dem Auftreten des transienten Signals. Verursacht wird dies durch Verschmieren des Zeitsignals über die Zeitspanne des für einen Rahmen nötigen Reihe von Abtastwerten. Je nach Anzahl der Abtastwerte (Transformationslänge) ist die Zeitauflösung entsprechend gering. Der Grund warum in der Regel kein Echo nach Auftreten des Ereignisses hörbar wird, liegt in der in Kapitel 3 beschriebenen im Vergleich zur Vorverdeckung deutlich stärker ausgeprägten Nachverdeckung. Das heißt ein mögliches (post) Echo wird einfach durch das Ereignis maskiert. Die Anwendung des „pre-echo Control“ Verfahrens welches das Maximum von Verdeckungsschwellwerten über den aktuellen, den letzten und den vorletzten Rahmen bildet, ergibt zu hohe Schwellwerte mit den im Standard angegebenen Koeffizienten. Der MPEG1 Layer 3 Standard sieht an dieser Stelle keine Rücktransformation der Energien aus dem kritischen Bandbereich in den Frequenzbereich vor. Möglicherweise sind die Koeffizienten auf Basis der eigentlich nachfolgenden Schritte (Transformation zu sogenannten „scalefactor bands“) angepasst. Aufgrund der Weiterverwendung der Verdeckungschwellwerte für eine Gewichtungsregel und der durch Transformation in den kritischen Bandbereich einhergehende Glättung der „zu einem Band gebündelten Frequenzlinien“ ist der Effekt des vorzeitigen Echos eher gering, zudem hat eine Anpassung der Koeffizienten bei informellen Hörtests zu schlechteren Ergebnissen geführt. Des Weiteren liegt der Fokus der Störgeräuschreduktion auf der Verarbeitung von Sprachsignalen, bei denen stark transiente Vorgänge und damit auftretende vorzeitige Echos eher selten sind. Der Standard wurde für Musik entwickelt, wo eine Kompenssation solcher Phänomene durchaus angemessen ist. Im Allgemeinen kann „pre-echo“ verringert werden, in dem die Transformationslänge der Charakteristik des Signals angepasst wird. Für transiente Vorgänge wird die Transformationslänge kleiner gewählt. Diesen Vorgängen wird über eine parallele Berechnung der Verdeckungsschwellwerte mit kurzer und langer Transformationslänge Rechnung getragen (siehe Abschnitt 4.2). Das „pre-echo control“ wird daher bei der Bestimmung der Verdeckungsschwellwerte ausgelassen.

4.1.2.3 Rücktransformation der Verdeckungsschwellwerte

Nach Anwenden der absoluten Hörschwelle (im Standard als Punkt k) gekennzeichnet, werden die Verdeckungsschwellwerte nicht wie im Standard vorgesehen in sogenannte „scalefactor bands“ transformiert oder daraus Signal zu (wie in Punkt n) beschrieben) Maskierungsverhältnisse (SMR) berechnet. Stattdessen wird die Rücktransformation in den Frequenzbereich wie für den Vorgänger MPEG1 Layer II (entsprechend des Punkts k) im psychoakustischen Modell 2 des Standards) vorgesehen, durchgeführt.

4.2 Modell 2: erweitertes Modell MP3GADV

4.2.1 Modellbeschreibung

Das erweiterte MPEG1 Layer 3 Modell, welches auf dem oben vorgestellten angepassten Modell von Gustafsson (Abschnitt 4.1) bezogen auf den Standard vervollständigt und erweitert, ist in der unteren Abbildung 4.1 zu sehen. Die Erweiterungen sind in rot gekennzeichnet. Die grobe Struktur des Modells gleicht dem Referenzmodell.

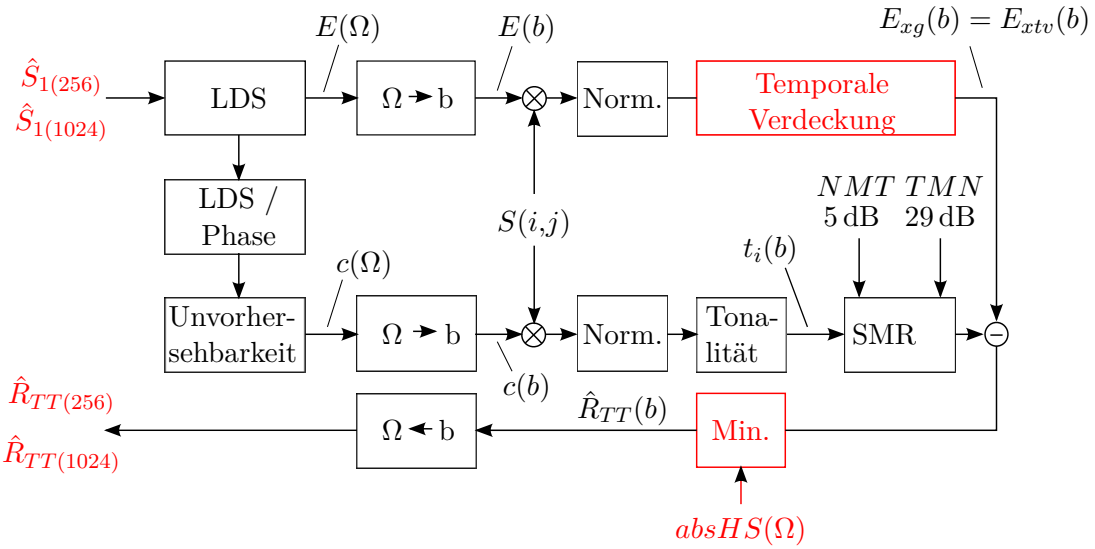


Abbildung 4.2: Eweitertes psychoakustisches Modell des MPEG1 Layer 3 Standards MP3GADV

4.2.2 Erweiterungen

4.2.2.1 Berücksichtigung der absoluten Hörschwelle

Zunächst wird die absolute Hörschwelle entsprechend der Abbildung 3.4) in Kapitel 3 hinzugefügt. Die in den Tabellen des Standards beschriebene absolute Hörschwelle stellt eine sehr grobstufige Quantisierung dieser absoluten Hörschwelle nach Zwicker dar und liefert entsprechend schlechtere Ergebnisse. Die Funktion der absoluten Hörschwelle nach Zwicker wird nach Rücktransformation der Verdeckungsschwellwerte in den Frequenzbereich entsprechend der unteren Gleichung angewendet.

$$\hat{R}_{TT}(\Omega) = \max(\hat{R}_{TT}(\Omega), absHS(\Omega)) \tag{4.1}$$

Der Variablenname $\hat{R}_{TT}(\Omega)$ steht hier stellvertretend für die im nächsten Unterabschnitt vorgestellten Verdeckungsschwellwerte unterschiedlicher Transformationslänge. Da die Hörschwelle eine absolute Größe ist, muss das Eingangsspektrum $\hat{S}_1$ skaliert werden, da Signale im verwendeten MATLAB Framework generell auf einen maximalen Wert von ± 1 normiert sind und dies nicht den Schalldruckpegel repräsentiert. In der oben erwähnten Abbildung der absoluten Hörschwelle ist auch der Schalldruckpegelbereich eines Sprachsignals eines einzelnen Sprechers eingezeichnet. Als Mittelwert kann man 45 bis 50 dB

ablesen. Im Standard ist jedoch angegeben, dass die Funktion der absoluten Hörschwelle auf das Maximum der Energie eines mit 16 bit kodierten Signals referenziert ist. Dies entspricht 96 dB, welche von der Funktion (in dB angegeben) subtrahiert werden und diese damit auf die Signalamplitude angepasst ist.

4.2.2.2 Zweifache Berechnung der Schwellwerte

Im Standard ist die Berechnung des Tonalitätsindexes auf Basis von Spektren unterschiedlicher Transformationslänge in Abhängigkeit der Frequenz vorgesehen. Das zur Berechnung des Tonalitätsindexes nötige Unvorhersehbarkeitsmaß $c(\Omega)$ (nach [5]) soll entsprechend des Standards für niedrige Frequenzen unter Verwendung des Leistungsdichtespektrums des Spachsignals mit 256 Abtastwerten und für höhere Frequenzen bis ca. 8 kHz mit einem Leistungsdichtespektrum mit 1024 Abtastwerten berechnet werden. Für die Frequenzen darüber wird das Unvorhersehbarkeitsmaß auf einen konstanten Wert gesetzt. Aufgrund dieser Beschreibung und des Erscheinungsdatums des Standards, wird davon ausgegangen, dass die Berechnung mangels verfügbarer Rechenleistung vereinfacht wurde. Zudem ist aus praktischen Gründen der Weiterverwendung des Tonalitätsindexes eine parallele Berechnung der Verdeckungsschwellwerte unter Verwendung unterschiedlich langer Spektren des geschätzten Sprachsignals sinnvoll. Darüber hinaus gewinnt man für eine lange Abtastzeit (1024 Abtastwerte) bei der diskreten Fouriertransformation eine höhere Frequenzauflösung und für eine kürzere Abtastzeit (256 Abtastwerte) eine höhere Zeitauflösung. Das in der Abbildung 4.2 dargestellte Modell wird also zweimal durchlaufen. Dazu werden jeweils ein Spektrum des geschätzten Sprachsignals benötigt, welche aus der ebenso zweifach durchgeführten konventionellen Störgeräuschreduktion der ersten Stufe ermittelt werden. Die diskrete Fouriertransformation wird jeweils mit den oben genannten Werten durchgeführt. Die Rahmenlängen betragen 768 und 192. Dies entspricht 75 % der Transformationslänge, was zu besseren Ergebnissen führt als bei 50 %. Der Rahmenvorschub wird so eingestellt, dass sich die Rahmen zur Hälfte überlappen. Anzumerken ist noch, dass für die Transformation des Spektrums der kurzen Transformationslänge von 256 Abstastwerten im Gegensatz zur langen nur die im Standard vorgesehenen 38 statt 61 Bänder bei einer Abtastfrequenz von 48 kHz verwendet werden.

4.2.2.3 Kombination der Verdeckungsschwellwerte

Die berechneten Mithörschwellen $\hat{R}_{TT(256)}$ und $\hat{R}_{TT(1024)}$ können auf verschiedene Weise kombiniert werden, um die für das Filtergewicht H_{psycho} nötige Mithörschwelle $\hat{R}_{TT}$ zu erhalten. Abbildung 4.3 zeigt schematisch die Berechnung für diesen. Da der berechnete Schwellwert $\hat{R}_{TT(256)}$ aus der kurzen Transformationslänge 4 Rahmen für die gleiche Zeitspanne des zum Spektrum $\hat{S}_{1(1024)}$ korrespondierende Zeitsignals $\hat{s}_1$ hat und $\hat{R}_{TT(1024)}$ dafür nur einen Rahmen aufweist, muss dieser auf die lange Transformationslänge umgerechnet werden.

Es gibt verschiedene Möglichkeiten den in der Abbildung mit „Berechnungsvorschrift 1“ gekennzeichneten Block zu gestalten: Da die höhere Zeitauflösung die zusätzliche Berechnung der Mithörschwelle mit kurzer Transformationslänge motiviert, liegt es nahe das Maximum oder Minimum jeder einzelnen Frequenzlinie über die 4 Rahmen zu bestimmen. Weist das geschätzte Sprachsignal $\hat{s}_1$ im Zeitbereich innerhalb der Zeitspanne eines Rah-

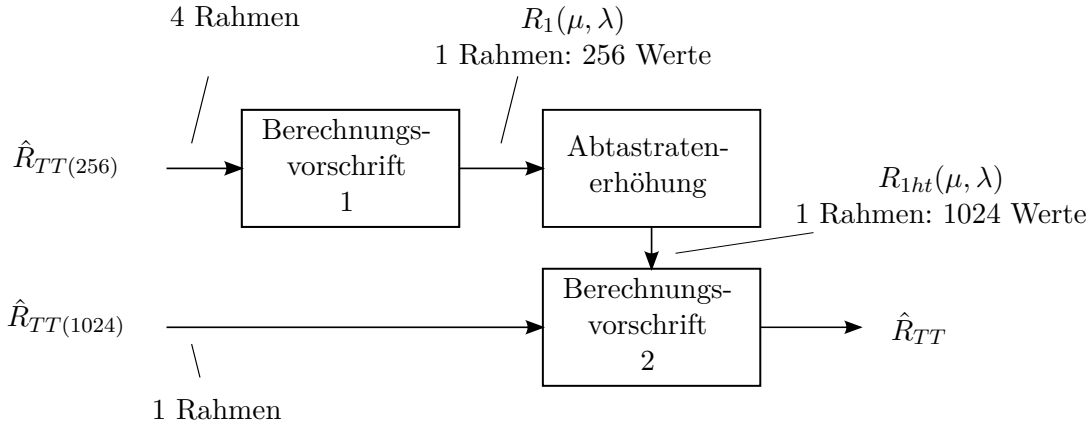


Abbildung 4.3: Das menschliche Gehör als System

mens der langen Transformationslänge ein kurzzeitiges Minimum oder Maximum auf, so ist dies möglicherweise nicht ausreichend im Spektrum mit hoher Transformationslänge (hier 1024 Abtastwerte) repräsentiert. Unter Annahme, dass die Schwellwerte mittels des Sprachsignalspektrums mit Transformationslänge betstimmt werden, führt dies dann zu Fehlern bei der Berechnung der geschätzten Mithörschwellen. Für den Fall das ein kurzzeitiges Maximum "unbemerkt„bleibt, wäre die berechnete Mithörschwelle zu klein. Da alle in dieser Arbeit behandelten Filterregeln den Verdeckungsschwellwert (oder das sich korrelierend verhaltene Erregungsmuster) und Leistungsdichtespektrum des geschätzten Rauschens ins Verhältnis setzen (Kapitel 5), fiele das resultierende Filtergewicht H_{psycho} der zweiten Stufe zu klein aus. Das verrauschte Signal würde auf Kosten des Sprachanteils stärker gedämpft als es die Maskierung zuließe. Der Fall des kurzzeitig auftretenden Minimums, welches aufgrund geringer Zeitauflösung der hohen Transformationslänge nicht im Spektrum ausreichend berücksichtigt ist, führt entsprechend auf für den Zeitpunkt des Auftretens des Minimums auf zu hohe Verdeckungsschwellwerte. Die Rauschunterdrückung des resultierenden Filtergewichts ist dann nicht ausreichend. Gegebenenfalls wird vorhandenes Musical Noise nicht unterdrückt. Der Fall der zu geringeren Rauschunterdrückung ist möglicherweise in der Wahrnehmung des Klangergebnis des verbesserten Sprachsignals schlechter als eine etwas zu starke Rauschunterdrückung und damit einhergehend hohe Dämpfung des Sprachanteils. Für die Berechnungsvorschrift 1 wird aufgrund besserer Ergebnisse bei der Störgeräuschreduktion das Minimum jeder einzelnen Frequenzlinie aus den 4 Rahmen ausgewählt. Die sich ergebenden Spektralwerte B1 für die 256 Frequenzlinien des betrachteten Rahmens λ_{256} können mit folgender Gleichung berechnet werden. Während der Index λ_{256} die Rahmen des Schwellwerts $\hat{R}_{TT(256)}$ angibt, bezeichnet $\lambda_{1024} = \lambda$ die Rahmen des Schwellwerts $\hat{R}_{TT(1024)}$, welcher dem Rahmenindex für das zu verarbeitende Signal entspricht.

$$R_1(\mu, \lambda) = \min\left(\hat{R}_{TT(256)}(\mu, \lambda_{256}), \ldots, \hat{R}_{TT(256)}(\mu, \lambda_{256} + 3\right) \tag{4.2}$$

Im Block „Abtastratenerhöhung“ wird R_1 auf 1024 Werte hochgetastet. Die resultierende Größe $R_{1ht}(\mu, \lambda)$ und der Verdeckungsschwellwert $\hat{R}_{TT(1024)}(\mu, \lambda)$ werden mittels Berechnungsvorschrift 2 kombiniert. Auch hier wird das Minimum jeder einzelnen Frequenzlinien beider Größen genommen. Wird statt dessen das Maximum genommen, verringert sich

der Vorteil der zweifachen Schwellertberechnung auf Basis unterschiedlich langer Transformationslängen. Der kombinierte Verdeckungsschwellwert $\hat{R}(\mu,\lambda)$ ist durch Gleichung 4.3 angegeben:

$$\hat{R}(\mu,\lambda) = \min\left(R_{1ht}(\mu,\lambda), \hat{R}_{TT(1024)}(\mu,\lambda)\right) \tag{4.3}$$

Alternativ kann die psychoakustische Entropie [32] als Berechnungsvorschrift 2 gewählt werden. Abhängig von der Struktur des Signals (eher transient oder stationär) wird der hochgetasteten Wert der kurzen Transformationslänge $R_{1ht}(\mu,\lambda)$ für transientere Signale und der Verdeckugschwellwert der hohen Transformationslänge $\hat{R}_{TT(1024)}(\mu,\lambda)$ für eher stationäre Signale verwendet. Die auf den betrachteten Rahmen λ bezogene psychoakustische Entropie pe stellt das Verhältnis von Verdeckungsschwellert $\hat{R}_{TT(1024)}(b)$ (der hohen Transformationslänge) im zu der Energie $E(b)$ kritischen Bandbereich dar:

$$pe = -\sum_{b=0}^{Z}\left(cbw_z(b)\cdot\frac{\hat{R}_{TT(1024)}(b)}{E(b)+1}\right) \tag{4.4}$$

Dabei ist $cbw_z(b)$ die im vorigen Kapitel vorgestellte kritische Bandbreite nach Zwicker in Bark. Z bezeichnet die Anzahl der Bänder. Die Verhältnisse des jeweiligen Bandes gehen für die breiteren Bänder stärker in die Summe ein. Der Betrag der psychoakustischen Entropie ist für transiente größer als für stationäre Signale. Die Entscheidung den Verdeckungschwellert $R_{1ht}(\mu,\lambda)$ zu nehmen, basiert auf einem definierten Schwellwert. Dieser wird anders als im Standard vorgesehen auf einen Wert von 200 angepasst. Die psychoakustische Entropie ergibt für transiente Signale geringfügig bessere Ergebnisse bzgl. der Sprachdämpfung. Allerdings nimmt der verbleibende Störgeräuschanteil, das wahrnehmbare „Musical Noise“, zu. Es treten leichte Nachhalleffekte auf.

4.2.2.4 Temporale Verdeckung

Zur Erweiterung des psychoakustischen Modells um eine Komponente, welche die temporale Verdeckung des menschlichen Gehörs einbezieht, wird das im FFT Modell des PEAQ Standards beschriebene Verfahren [37] (Abschnitt: Time domain spreading) verwendet. Die Gleichungen sollen hier zum besseren Verständnis kurz erwähnt werden. Die temporal verschmierte Energie E'_{xtv} berechnet sich aus der Erregung E_x wie folgt.

$$E'_{xtv}(b,\lambda) = a\cdot E'_{xtv}(b,\lambda-1) + (1-a)\cdot E_x(b,\lambda) \tag{4.5}$$

Es findet also eine Tiefpassfilterung über zwei Rahmen statt. Daher ist nur die wesentlich wichtigere Nachverdeckung repräsentiert. Vorverdeckung wird nicht berücksichtigt. Der Koeffizient a ist durch Gleichung 4.6 beschrieben:

$$a = e^{-\frac{r_v}{f_s}\cdot\frac{1}{\tau}} \tag{4.6}$$

Der Exponent setzt sich aus dem Rahmenvorschub r_v und der Abtastfrequenz f_s zusammen. Da der Rahmenvorschub ein Bruchteil der Rahmenlänge ist, wird diese im Koeffizienten berücksichtigt und dieser daher auf das Verhältnis von Rahmenlänge und Abtastfrequenz angepasst. Für große Rahmenlängen ist der Rahmenvorschub größer. Bei gleichbleibender Abtastfrequenz resultiert ein kleinerer Koeffizient a. Der Einfluss der verschmierten

Energie E'_{xtv} des vorhergehenden Rahmens fällt dann geringer aus, da die Zeitspanne eines Rahmen größerer Länge höher ist und nun die Nachverdeckung zeitlich begrenzt ist. Umgekehrt wird bei kleineren Rahmen der Einfluss der verschmierten Energie des vorhergehenden Rahmens größer. Die Zeitkonstante τ ist neben der Mittenfrequenz f_c (in Hz) des betrachteten Bandes b unter anderem abhängig von den für dieses psychoakustische Modell angepassten Zeitkonstanten τ_{min} und τ_{100}:

$$\tau = \tau_{min} + \frac{100}{f_c(b)} \cdot (\tau_{100} - \tau_{min}) \tag{4.7}$$

Der Standard sieht für die Zeitkonstanten τ_{min} und τ_{100} 8 bzw. 30 ms vor. Die daraus resultierenden Werte der Zeitkonstanten τ reichen von ca. 36 ms bei einer Mittenfrequenz des Bandes von 80 Hz bis ca. 8 ms bei 15 kHz und für darüber liegende Frequenzen. Allerdings weist das zu verarbeitende Spektrum des geschätzten Sprachsignals bei diesem Standard eine Transformationslänge von 2048 Abtastwerten auf. Die Zeitkonstante liegt daher noch innerhalb der Zeitspanne eines Rahmens (43 ms). Für das hier betrachtete psychoakustische Modell 2 des MPEG1 Layer 3 Standards mit Transformationslängen von 1024 entsprechend 21 ms und 256 (5 ms) übersteigen die Zeitkonstanten teilweise die Zeitspanne eines Rahmens. Der kompensierende Effekt des Bruchs in Gleichung 4.6, welcher den Rahmenvorschub einbezieht, reicht nicht aus, um bei Anwendung der temporalen Verdeckung bei vergleichsweise kurzen Transformationslängen einen Nachhalleffekt zu verhindern. Die temporale Verdeckung wird daher nur auf die Erregung mit langer Transformationslänge angewendet. Informelle Hörtests bis zum Verschwinden des Nachhalls führen auf Werte von 25 ms und 6 ms für Zeitkonstanten τ_{100} und τ_{min}. Die Ursache liegt möglicherweise in dem vom PEAQ Standard abweichenden Verhältnis von Transformationslänge und Rahmenlänge (75 % statt 50 %).

Kompensation bei stark transienten Signalen Für transiente Signale sieht der Standard eine Kompensation vor, welche die durch Tiefpassfilterung der temporalen Verdeckung geglätteten Spitzen im Spektrum (Maxima) vor der Filterung bewahrt. Es wird damit verhindert, dass die aus den Erregungsmustern abgeleiteten Verdeckungsschwellwerte nicht zu klein bzgl. ihrer maximal nutzbaren Amplitude ausfallen. Die Kompensation berücksichtigt das Maximum der Werte der berechneten temporal verschmierten Energie E'_{xtv} und der Erregung E_x. Ist letztere aufgrund eines transienten Vorgangs größer, bleibt diese damit verknüpfte hohe Energie für die Berechnung des Verdeckungsschwellwerts erhalten. Die Kompensation scheint angemessen zu sein, wenn man bedenkt, dass dieses Modell zur temporalen Verdeckung Teil des PEAQ Standards ist, der eine Transformationslänge von 2048 (entsprechend 43 ms) Abtastwerten hat und daher eine im Vergleich zum psychoakustischen Modell 2 des MPEG1 Standards geringe Zeitauflösung aufweist. Die betrachtete Zeitspanne liegt bei letzterem für einen Rahmen bei ca. 21 ms. Daher scheint eine Kompensation weniger nötig. In informellen Hörtests hat sich die Kompensation sogar als Verschlechterung erwiesen und wird daher weggelassen.

Energieerhaltung Nach Anwedung der temporalen Verschmierung ergibt sich wie bei der spektralen Spreizung eine veränderte Gesamtenergie, da Energieanteile des vorhergehenden Rahmens auf die gewichete Energie des nachfolgenden Rahmens addiert werden

(Gleichung 4.5). Um Energieerhaltung zu bewahren, wird das temporal verschmierte Spektrum jeden Rahmens $E_{xtv'}(\lambda)$ mit einem Normierungsfaktor n_{tv} multipliziert, der sich wie folgt berechnet:

$$n_{tv} = \sum_{\lambda} \left(\frac{E_x(b,\lambda)}{E'_{xtv}(b,\lambda)} \right) \tag{4.8}$$

Die Energien vor und nach Anwendung der temporalen Verdeckung werden jeweils über alle Rahmen summiert und die daraus resultierenden Gesamtenergien ins Verhältnis gesetzt. Vor Einführen der Reskalierung der verschmierten Energien konnte man bei informellen Hörtests Nachhalleffekte vernehmen. Dies ist auf zu große Verdeckungsschwellwert aufgrund zu hoch berechneter Energien $E_{xtv'}(b,\lambda)$ zurückzuführen. Durch die Reskalierung wird dieser Effekt unterbunden. Nach Multiplikation erhält man die Energie $E_{xtv}(b,\lambda)$

$$E_{xtv}(b,\lambda) = n_{tv} \cdot E'_{xtv}(b,\lambda) \tag{4.9}$$

4.3 Modell 3: FFT Modell des PEAQ Standards PEAQFFT

Das im folgenden vorgestellte dritte Modell wird gewählt, um fehlende Aspekte bei dem originalen psychoakustischen Modell 2 des MPEG1 Layer 3 Standards bzgl. der am Ende des Kapitels 3 erläuterten Anforderungen zufrieden zu stellen.

4.3.1 Modellbeschreibung

In der unteren Abbildung 4.4 ist das Blockdiagramm des FFT Modells des PEAQ Standards [1] und [37] dargestellt. Die grün markierten Blöcke und Größen kennzeichnen die Unterschiede zum originalen MPEG1 Layer 3 Modell (ohne Erweiterungen). Die rot markierten Bereiche stellen nötige Anpassungen und Erweiterungen dar, welche im nächsten Abschnitt erläuter werden.

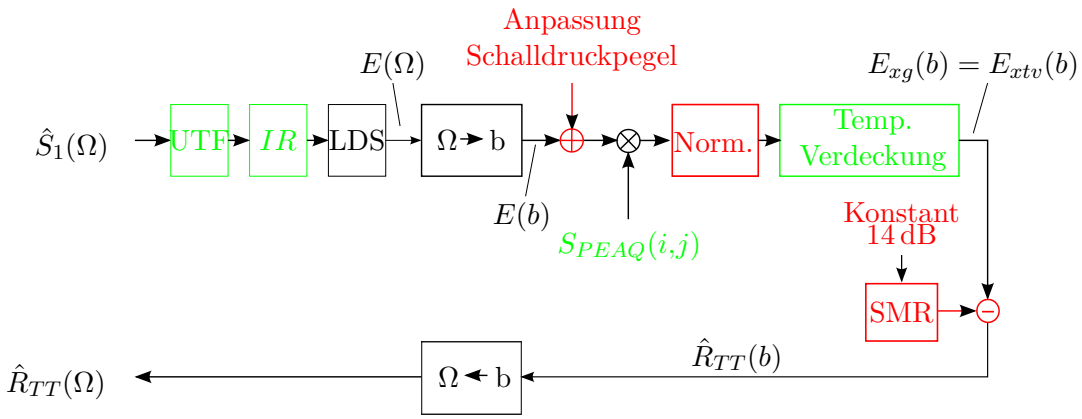


Abbildung 4.4: Blockdiagramm des psychoakustischen Modells PEAQFFT

Die grobe Struktur gleicht den vorhergehenden Modellen und orientiert sich an den in der Einleitung vorgestellten Verfahrensschritten. Zunächst findet eine Gewichtung des Spektrums des geschätzten Sprachanteils $\hat{S}_1(\Omega)$ mittels der Übertragungsfunktion des

Außen- und Mittelohres statt. Die Übertragungsfunktion UTF ist in Abbildung 3.2 in Kapitel 3 aufgetragen. Dem gefilterten Spektrum wird das interne Rauschen (Abbildung 3.3) hinzugefügt. Die Transformation, der aus dem resultierenden Spektrum berechneten Energien in den Bereich der kritischen Bänder, erfolgt mit der Transformationsvorschrift nach Schroeder [28]. Die zugehörige Gleichung 3.5 und Abbildung 3.5a sind im vorhergehenden Kapitel angegeben. Dabei existieren zwei Varianten: eine mit einem Bandabstand von $dz = 0.25$ Bark. Dies führt auf 109 Bänder für das vom Modell verarbeitete Spektrum von 80 bis 18000 Hz. Die zweite Variante nutzt nur einen Abstand von $dz = 0.5$ Bark mit 59 Bändern. Bei beiden Varianten erfolgt die Gruppierung von Frequenzlinien zu einem Band (Frequenzgruppe). Im Vergleich zum MPEG1 Layer 3 Modell wird jedoch die Frequenzauflösung, d.h. die Bandbreite, welche durch eine Linie repräsentiert ist, berücksichtigt. Liegt die Bandbreite auf der Grenze zwischen zwei Bändern, das heißt ragt zu einem Teil in das untere zum anderen Teil in das nächst höhere Band, so wird die jeweilige Energie entsprechend der Anteile am jeweiligen Band aufgeteilt.

Nach der Transformation erfolgt eine im Standard nicht vorgesehene Anpassung des logarithmischen Schalldruckpegels durch Addition eines Abstands „Offset“. Als wesentlicher Unterschied zu den vorhergehenden Modellen ist die Faltung mit einer von der Frequenz und Amplitude der Energie abhängigen Spreizfunktion zu erwähnen. Die Spreizfunktion des psychoakusitschen Modells 2 des MPEG1 Layer 3 Standard ist nur frequenzabhängig und berücksichtigt nicht die Amplitude. Anschließend erfolgt eine Normierung, welche angepasst wird. Die temporale Verdeckung ist im Standard schon gegeben, berücksichtigt allerdings nur die Nachverdeckung. Der Grund dafür liegt in der Anstiegszeit des Verdeckungsschwellwerts bei Vorverdeckung. Diese beträgt 2 bis 5 ms und mehr. Da die Zeitauflösung des FFT basierten Gehörmodels des PEAQ Standards ungefähr 20 ms beträgt, wird die Vorverdeckung nicht berücksichtigt [37].

Das vom Erregungsmuster zu subtrahierende Signal-Maskierungsverhätnis SMR erfolgt nach eigenen Verfahren, die unten beschrieben werden. Der resultierende Verdeckungsschwellwert wird mit der inversen Transformationsvorschrift (Gleichung 3.7 Kapitel 3) in den Frequenzbereich zurück transformiert. Es ist anzumerken, dass eine Tonalitätsberechnung wie beim psychoakustischen Modell 2 des MPEG1 Layer 3 Standards zur Charakterisierung des Spektrums und darauf basierendem SMR im Standard nicht vorgesehen ist.

4.3.2 Erweiterungen

In diesem Unterabschnitt sollen einige Anpassungen und Erweiterungen vorgestellt werden, die nötig sind, um möglichst gute Ergebnisse bei Verwendung des Modells in der Störgeräuschreduktion zu erzielen.

4.3.2.1 Skalierung des internen Rauschens

Das interne Rauschen IR ist ganuso wie die absolute Hörschwelle eine absolute Größe. Da das gefilterte Spektrum aus der Transformation des auf den Betrag 1 (0 dB) normierten geschätzten Sprachsignals hervorgeht, muss das interne Rauschen auf dieses skaliert werden. Das Leistungsdichtespektrum des internen Rauschens ist als Funktion der Frequenz in Dezibel angegeben und auf ein 16 bit Signal referenziert. Daraus ergibt sich eine maxi-

male Energie eines Sinustons von ca. 96 dB. Diese 96 dB werden vom internen Rauschen abgezogen, womit es der Skalierung der in der Simulationsumgebung verwendeten Signale entspricht.

4.3.2.2 Anpassung des Schalldruckpegels für die Spreizfunktion

Die obere Flanke der Spreizfunktion $S_{peaq,\lambda}(i,j)$ ist amplitudenabhängig. Wie oben beschrieben repräsentiert das Spektrum $\hat{S}_1$ nicht den tatsächlichen Schalldruckpegel. Die Energie im Bandbereich $E(b)$ wird auf Basis des Spektrums berechnet. Die Skalierung muss daher für die Spreizung angepasst werden. Im Kapitel 3 ist in Abbildung 3.4 der Bereich für Sprache typischen Schalldruckpegels angegeben. Man kann dort 45 dB als groben Mittelwert ablesen, dieser Wert wird auch als zeitlicher Mittelwert für einen einzelnen Sprecher bei normaler Sprechlautstärke angegeben [14].

Zur Veranschaulichung ist die Gleichung für die obere Flanke s_{oF} unten angegeben:

$$s_{oF} = -24 - \frac{230\,\mathrm{Hz}}{f_c} + 0{,}2 \cdot L \tag{4.10}$$

Die Flanke weist die Einheit $dB/Bark$ auf. Die Mittenfrequenz des betrachteten Bands ist durch die Variable f_c repräsentiert. L ist das lokale Energieniveau, welches mit dem Schalldruckpegel L_p korrelliert. Die Flanke ist für die unteren Bänder stärker von dessen Mittenfrequenz f_c abhängig. Daher ist der Einfluss durch das lokale Energieniveau L hier noch relativ gering. Zu höheren Bändern hin verschwindet der Term, welcher die Mittenfrequenz enthält und der Einfluss der lokalen Energie L steigt an. Daher ist gerade für die höheren Bänder eine korrekte Skalierung sehr wichtig, um die Verläufe der von Zwicker gemessenen Mithörschwellen (Abbildung 3.9) möglichst gut zu modellieren. Obwohl das lokale Energieniveau in Gleichung 4.10 aufgrund der Filterung mittels der Übertragungsfunktion des Außen- und Mittelohres (Abbildung 3.2 in Kapitel 3) nicht direkt dem Schalldruckpegel entspricht, besteht dennoch eine starke Korrelation. Die Übertragungsfunktion weist für den Frequenzbereich der Sprache relativ niedrige Dämpfungswerte auf. Das heißt das Spektrum wird in diesem Bereich nicht so stark verändert - zumindest nicht bzgl. des Dezibel Werts. Die Transformation hat keinen Einfluss, da Energien einzelner Frequenzlinien addiert werden. Die Addition eines logarithmischen Abstands entspricht einer skalaren Multipikation. Informelle Hörtests bestätigen den theoretischen Wert von 45 dB des zu addierenden Abstands. Die Energien werden bei der nachfolgenden Normierung (nächster Unterabschnitt) wieder auf das vorherige Niveau skaliert.

4.3.2.3 Normierung nach Faltung mit Spreizfunktion

Die Spreizfunktion $S_{peaq,\lambda}(i,j)$ des PEAQ Standards ist aufgrund der Amplitudenabhnängigkeit für jeden Rahmen unterschiedlich. Die Indizes i und j kennzeichnen den Maskierer bzw. das zu maskierende Band. Der Vorgang der Faltung mit der Spreizfunktion kann prinzipiell in zwei Schritte aufgeteilt werden. Als Erstes werden die einzelnen Energien gespreizt und im zweiten Schritt superponiert. Die Fläche unterhalb der Spreizfunktion variiert in Abhängigkeit der Mittenfrequenz und des Schalldruckpegels. Der Standard schreibt eine Normierung vor, welche die Fläche unterhalb der Spreizfunktion für alle Mittenfrequenzen und ein konstantes Energieniveau L von 0 dB auf den Wert 1 setzt. Somit ist

die Spreizfunktion $S_{peaq,\lambda}(i,j)$ bereits normiert. Da insgesamt bei hohen lokalen Energieniveaus L die Spreizung zu hohen Frequenzen stärker ausgeprägt ist, führt dies allerdings zu Flächen größer 1 unterhalb der entsprechenden Spreizfunktion. Daraus resultiert gerade bei Auftreten von hohen Schalldruckpegeln im Signal (bzw. hohen lokalen Energieniveaus) eine zusätzliche Energie im Vergleich zur Energie des dem Modell zugeführten Spektrums $\hat{S}_1$. Zusätzlich wird die Gesamtenergie durch die nicht lineare Superposition der gespreizten und skalierten Energien E'_x verändert. Mathematisch ist der Vorgang in der unteren Gleichung beschrieben:

$$E_{xg}(\lambda) = [S_{peaq,\lambda}(i,j)^a \cdot E(\lambda)^a]^{1/a} \tag{4.11}$$

Die normierte und mit dem Faktor a potenzierten Spreizfunktion $S_{spread,\lambda}(i,j)$ ist eine Matrix und wird mit der skalierten und ebenfalls mit dem Faktor $a = 0{,}4$ potenzierten Energie $E'(b)$ multipliziert. Das Ergebnis wird mit dem Kehrwert von a potenziert. Wegen der Realisierung der Faltung als Matrix wurde die Potenz a auf die Faktoren einzeln angewandt. Die Superposition entspricht dem in Kapitel 3 vorgestellten Gesetz (Gleichung 3.14) nach [24]), welches durch Untersuchungen den Wert für den Faktor a festgelegt hat.

Zur Berechnung der Gesamtenergie vor der Faltung mit der Spreizfunktion werden die Rahmen des quadrierten Betragsspektrums $|\hat{S}_1(\Omega,\lambda)|^2$ summiert. Für die resultierde Gesamtenergie nach Anwendung der Faltung wird die Summe der Rahmen über den gespreizten Energien (Erregungsmustern) E_{xg} gebildet. Der Vergleich der Gesamtenergien führt auf die Verletzung des Energierhaltungssatzes.

Die Energieerhaltung wird von dem Standard nicht berücksichtigt. Allerdings werden die Erregungsmuster dort nicht zur Störgeräuschreduktion verwendet, sondern sind für den Zweck, Evaluationsparamter zu bestimmen, skaliert. Es wird eine Normierung entsprechend der Energieerhaltung durchgeführt. Der Normierungsfaktor n_{sp} ergibt sich wie folgt.

$$n_{sp} = \sum_{\lambda} \left(\frac{\hat{S}_1(\Omega,\lambda)}{E_{xg}(\Omega,\lambda)} \right) \tag{4.12}$$

Anschließend werden alle Rahmen $E_{xg}(\Omega,\lambda)$ mit dem Normierungsfaktor multipliziert.

4.3.2.4 Normierung nach Anwenden der temporalen Verdeckung

Die Normierung nach Anwenden der temporalen Verdeckung verläuft analog zu der in Unterabschnitt 4.2.2.3 geschilderten Normierung. Jedoch wird der Normierungsfaktor auf die Energien $E_{xtv}(b,\lambda)$ angewendet. Diese sind das Ergebnis der Gleichung 4.16, welche eine Kompensation von transienten Vorgängen beschreibt und Teil der temporalen Verdeckung des PEAQ Standards ist.

$$E_{xtv}(b,\lambda) = \max\left([E'_{xtv}(b,\lambda), E_x(b,\lambda)\right] \tag{4.13}$$

4.3.2.5 Signal zu Maskierungsabstand

Die im Standard vorgesehene Berechnung des Maskierungsabstands führt zu zu geringen Werten (maximal 7 dB). Die Verdeckungsschwellwerte, welche aus der Subtraktion des Signal zu Maskierungsabstands vom Erregungsmuster E_x berechnet werden, fallen dann

zu hoch aus. Dies führt bei allen Filterregeln zu einer geringen Störgeräuschreduktion, da diese das Leistungsdichtespektrum der Verdeckungsschwellwerte mit dem geschätzten Leistungsdichtespektrum des Störanteils ins Verhältnis setzen. Daher werden andere Berechnungen des Signal zu Maskierungsabstands entwickelt.

Konstanter Signal zu Maskierungsabstand Zunächst wird ein konstanter Signal zu Maskierungsabstand gewählt. Dieser wird je nach angewendeter Filterregel auf 14 bis 21 dB eingestellt. Die Werte ergeben sich aus Überlegungen von Thiede [38], welcher für Ton zu den Rausch Maskierungsabstand Gleichung 3.15 (Kapitel 3) angibt. Der Abstand startet bei 15 dB für das niedrigste Band und steigt dann pro Bark zu höheren Bändern um 1 dB an. Die Vereinfachung auf einen konstanten Wert liefert mindestens genauso gute Ergebnisse. Dies liegt möglicherweise in der Tatsache, dass das verrauschte Signal ohnehin einer Tiefpassfilterung von 16 kHz unterzogen wird und ohnehin die Energieanteile in den oberen Bändern sehr gering sind, so dass schon die unteren Grenzen der Filtergewichte wirken. Ein niedrigerer Verdeckungsschwellwert aufgrund eines höheren Signal zu Maskierungsabstandes hat also keinen Einfluss auf das Filtergewicht mehr.

Tonalitätsbasierter Signal zu Maskierungsabstand Es liegt nahe die Tonalität ähnlich dem MPEG1 Layer 3 Standard zu bestimmen und für die Berechnung des Signal zu Maskierungsabstandes zu benutzen. Es werden zwei Möglichkeiten erprobt: Die Tonalitätsberechnung wird einfach aus dem MPEG1 Layer 3 Standard adoptiert. Als weitere Alternative wird ein Tonaltitätsmaß auf Basis des spektralen Flachheitsmaßes (nach [5]) ermittelt.

Tonalitätsindex nach MPEG1 Layer 3 Trotz Anpassungen der Konstanten in den Gleichungen zur Bestimmung des Tonalitätsindexes $t_i(b)$ liefert die Berechnung [3] für diesen stets zu hohe Werte (nahe 1). Das heißt das Spektrum wird durch den Index sehr breitbandig (flach) eingeschätzt. Zur Erinnerung sei noch einmal darauf hingewiesen, dass der Tonalitätsindex für einen Ton den Wert 1 annimmt, und für ein sehr breitbandiges Signal (flaches Spektrum) den Wer 0. Ist der Wert des Indexes zu hoch, dann fällt der sich ergebende Signal zu Maskierungsabstand SMR entsprechend folgender Gleichung relativ hoch aus (Unterpunkt h) [19].

$$SMR(b) = t_i(b) \cdot TMN + (1 - t_i(b)) \cdot NMT \tag{4.14}$$

Die Größen TMN und NMT wurden im Kapitel 3 vorgestellt und betragen im Layer 3 29 und 6 dB. Der zu große SMR führt folglich auf einen zu niedrigen Verdeckungsschwellwert. Die Verdeckungschwellwerte ergeben sich, wie in den Blockdiagrammen oben zu sehen, aus der Subtraktion (in Dezibel) des Signal zu Maskierungsabstand SMR vom Erregungsmuster E_{xg}. Daraus ergibt sich für alle im nächsten Kapitel vorgestellten Filterregeln ein zu niedriges Filtergewicht mit entsprechend überhöhter Sprachdämpfung. Es wird versucht, mittels niedriger Werte für den Ton zu Rauschen Maskierungsabstand TMN, Abhilfe zu schaffen. Dabei wird die bandabhängige Funktion $TMN(b)$ aus dem Layer 2 entnommen, welche für die sprachrelevanten Bänder Werte um die 24 dB aufweist. Der vorher kostante Wert in der Gleichung wird durch diese Funktion ersetzt. Bei dieser Variante entfällt der

[3] Im Standard sind die Gleichungen als Unterpunkte c) bis g) des psychoakustischen Modells 2 zu finden

zweite Summand in Gleichung 4.14. Dies führt neben dem kleineren Wert für TMN zusätzlich zu kleineren Signal zu Maskierungsabständen. Auch bei weiterer Anpassung der Werte verbessern sich die Ergebnisse nicht. Es lassen sich auch keine Informationen über die Herleitung der zur Berechnung des Tonalitätsindexes $t_i(b)$ verwendeten Konstanten finden.

Tonalitätsbestimmung mittels des spektralen Flachheitsmaßes Alternativ kann die Tonalität auch mittels des spektralen Flachheitsmaßes SFM nach [5], [34] berechnet werden. Dies weist für tonale Signale hohe negative Dezibel Werte von ca. $-60\,\mathrm{dB}$ auf. Sprachsignale ergeben für den Frequenzbereich von 200 bis 3200 Hz einen Wert von $-20\,\mathrm{dB}$, breitbandige Signale bis zu $0\,\mathrm{dB}$ [5]. Das spektrale Flachheitsmaß wird auf das Teilspektrum der einem Band zugeordneten Frequenzlinien im Frequenzbereich berechnet. Für die niedrigere Bänder resultiert dies in tonalen Werten, da diese nur wenige Frequenzlinien enthalten. Ähnlich der Gleichung 4.14 wird auf Basis des sogenannten Tonalitätsmaßes α der Signal zu Maskierungsabstand (in dB) berechnet:

$$SMR(b) = \alpha(b) \cdot (14{,}5 + b)\,\mathrm{dB} + (1 - \alpha(b)) \cdot 5{,}5\,\mathrm{dB} \tag{4.15}$$

Der erste Faktor $(14{,}5 + b)$ ist der Gleichung 3.15 in Kapitel 3 nach Thiede [38] entlehnt. Zu beachten ist, dass b den Bandindex angibt und z in der erwähnten Gleichung die Tonheit in Bark. Der Wert von $5{,}5\,dB$ entspricht in etwa dem aus dem Layer 3 bekannten Rausch zu Ton Maskierabstand NMT. Das Tonalitätsmaß α ist durch das Verhältnis des spektralen Flachheitsmaßes zu dem Referenzwert des spektralen Flachheitsmaßes für einen Ton gegeben:

$$\alpha = min\left(\frac{SFM(b)}{-60\,dB}, 1\right) \tag{4.16}$$

Auch hier sind die Resulate deutlich schlechter als bei Verwendung des konstanten Signals zu Maskierungsabstands wie oben beschrieben. Der Grund liegt möglicherweise in der hohen Transformationslänge (2048 Abtastwerte) des FFT Modells des PEAQ Standards. Die Gleichungen zur Berechnung der Tonalität 4.14 und 4.15 sind vermutlich für kürzere Transformationslängen entwickelt worden – zumindest ist dies für erstere Gleichung aufgrund der deutlich kürzeren Transformationslänge beim MPEG1 Layer 3 Modell anzunehmen. Ein weiterer Grund lässt sich bei dem mit der hohen Transformationslänge verbundene Betrachtungszeitraum von knapp 43 ms, bzw. für einen Rahmen ca. 21 ms ausmachen. Die Zeitspanne ist für eine Tonalitätsbewertung möglicherweise zu lang. Bei dem MPEG1 Layer 3 Modell ist die Zeitspanne halb so groß. Das könnte auch das Weglassen eines tonalitätsbasierten Signal zu Maskierungsabstands im PEAQ Standards motiviert haben.

SNR gemittelt über Frequenzlinien innerhalb eines Rahmens Zuletzt soll der Versuch erläutert werden, mittels des Störabstands den Ton zu Rauschen Maskierungsabstand zu gewichten und das Ergebnis als Signal zu Maskierungsabstand zu betrachten. Die Idee liegt der Tatsache zu Grunde, dass bei niedrigen Störabständen verstärkt „Musical Noise“ auftritt und durch die Anpassung der Verdeckungsschwellwerte nach unten, dies zu einer stärkeren Dämpfung aufgrund eines kleineren Filtergewichts H_{psycho} führen würde. Umgekehrt würde bei hohen Störabständen der Signal zu Maskierungsmaskierungsabstand

klein, der Verdeckungsschwellwert und damit das Filtergewicht groß. Das Vorhaben zielt auf eine Verringerung der segmentellen Sprachdämpfung gegenüber der oben beschriebenen Methode ab, welche ein konstantes Signal zu Maskierungsverhältnis verwendet. Der geschätzte Störabstand $SNR(b)$ wird zunächst für jedes Band aus dem Verhältnis der Schätzung des Leistungsdichtespektrums des Sprachanteils und des Störanteils gebildet. Letztere werden aus der ersten Stufe der Störgeräuschreduktion gewonnen (vgl. Abbildung 2.5 in Kapitel 2. Dazu wird das Spektrum $\hat{R}_{nn}(b)$ in den Bereich der kritischen Bänder transformiert. Die Energie $E(b)$ erfolgt analog aus der Transformation der Energie des geschätzten Sprachanteils $\hat{S}_1$ entsprechend dem Blockdiagramm in Abbildung 4.4.

$$SNR(b) = \frac{E(b)^2}{\hat{R}_{nn}(b)} \tag{4.17}$$

Als allgemeine Berechnungsregel wird die folgende Gleichung verwendet:

$$SMR(b) = (1 - H_{SMR}(b)) \cdot TMN; \tag{4.18}$$

Dabei wird für $H_{SMR}(b)$ in einer ersten Variante der Störabstand $SNR(b)$ und einer zweiten Variante die Wiener Filterregel

$$H_{SMR}(b) = \frac{SNR(b)}{1 + SNR(b)} \tag{4.19}$$

eingesetzt. Das Ton zu Rauschen Verhältnis TMN wird unter Verwendung der Werte und Funktionen des Layer 2 und des Layer 3 [19] sowie der Werte, die auch schon bei der tonalitätsbasierten Berechungen genutzt werden, variiert. Diese Methode führt zu Verzerrungen und damit zu einer Verschlechterung des Klangergebnisses. Es wird angenommen, dass die Fehler bei Schätzung des Störabstands gerade bei niedrigen realen Störabständen besonders hoch sind und dadurch die Gewichtung des Ton zu Rauschenabtands fehlschlägt. Daraufhin wird für jedes Band eine Glättung über zwei Rahmen des geschätzten Störabstands $SNR(b)$ vollzogen, um mögliche Fehler zu mildern. Ein anderer Versuch verwendet den segmentellen Störabstand und führt damit auch zu einem innerhalb eines Rahmens konstanten Signal zu Maskierungsverhältnis SMR. Beide Versuche bringen keine Verbesserungen, sondern rufen zusätzliche Verzerrungen oder teilweise zu geringe oder zu hohe Sprachdämpfung hervor.

Abschließend kann gesagt werden, dass der konstante Signal zu Maskierungsabstand zu den besten Ergebnissen (siehe Kapitel 6) führt, was theoretisch in grober Näherung durch die Gleichung nach Thiede [38] (Gleichung 3.15 in Kapitel 3) gestützt wird.

4.4 Modell 4: Filterbank Modell des PEAQ Standards PEAQFB

Als viertes Modell soll das Filterbank Modell des PEAQ Standards [37] vorgestellt werden.

4.4.1 Modellbeschreibung

Die in der Einleitung erwähnten Verfahrensschritte treten auch bei dem Filterbank-Modell auf, jedoch in einer anderen Reihenfolge. Die Schritte sollen anhand des Blockdiagramms des Modells (Abbildung 4.5) kurz erläutert werden.

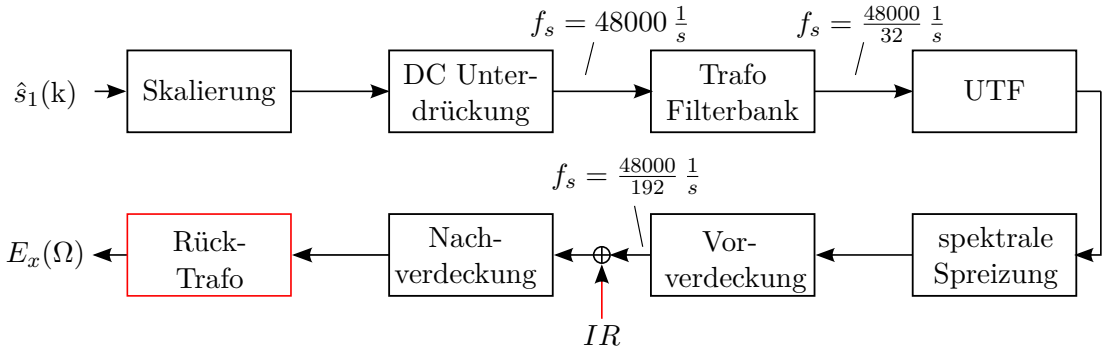


Abbildung 4.5: Blockschaltbild des PEAQ Filterbank Modells

Statt des Spektrums wie bei den FFT Modellen wird dem Filterbank-Modell das zeitdiskrete Signal $\hat{s}_1(k)$ zugeführt und skaliert. Anschließend erfolgt eine DC-Unterdrückung, bzw. Hochpassfilterung mittels IIR Filtern, da die Filterbank Gleichanteile nicht verarbeiten kann. Anschließend erfolgt eine Transformation des zeitdiskreten Signals in die 40 Bänder der Filterbank. Diese ergeben sich aus der in Kapitel 3 in Abbildung 3.6 gezeigte Äquivalentrechteckbandbreite (kurz ERB). Jedes Band ist durch ein Filterpaar repräsentiert, dessen Phasenantworten um 90 ° phasenverschoben sind. Der Ausgang des zweiten Filters stellt also den Imaginärteil des Ausgangs des ersten Filters dar. Das zeitdiskrete Signal wird also in ein analytisches Signal $s_a(k)$ umgerechnet, um die Einhüllende der Bandsignale zu berechnen. Dies ermöglicht die darauf folgende Unterabtastungen [36].

$$s_a(k) = \hat{s}_1(k) - j\hat{s}_1(k) \tag{4.20}$$

Nach Unterabtastung um den Faktor 32 wird das Signal in den Bändern mittels der gleichen Außen- und Mittelohrfunktion wie im PEAQ FFT Modell gefiltert. Die Unterabtastung führt zur Datenreduktion bei ausreichender Auflösung [38]. Die spektrale Spreizung erfolgt mit ähnlichen Flanken wie beim FFT Modell. Allerdings weist sie zusätzliche Eigenschaften auf: Zur Beibehaltung der Bandpasscharakteristik ist bei großen Amplituden des Bandpasssignals die Flanke auf −4 dB begrenzt. Der Verlauf der Spreizfunktion entspricht qualitativ dem Verlauf der dreieckförmigen Mithörschwelle in Abbildung 3.8 Kapitel 3. Das Gehör reagiert nicht direkt auf Schalldruckpegeländerungen, weist daher eine gewisse Latenz auf. Diesem Phänomen wird durch Glättung der oberen Flanke der Spreizfunktion mittels Tiefpassfilterung Rechnung getragen [37]. Die gespreizten Abtastwerte innerhalb der Bänder der jeweiligen Filterpaare werden nun in Energien umgerechnet. Dies entspricht der Betragsbildung des verarbeiteten analytischen Signals innerhalb jeden Bandes und wird als Gleichrichtung bezeichnet. Der Vorgang ist im Blockdiagramm dem Block „spektrale Spreizung“ zugeordnet. Da die Gleichrichtung erst nach der spektralen Spreizung durchgeführt wird, bleibt der Zusammenhang zwischen temporaler und spektraler Charakteristik der Filter erhalten [36]. Real- und Imaginärteil werden parallel gespreizt. Die Phase bleibt erhalten. Die Ausgangssignale entsprechen daher Filtern, welche direkt die durch die Spreizfunktion repräsentierten Flanken des Gehörs realisieren würden. Zusätzlich zu den vorhergehenden Modellen, weist dieses Modell als einziges eine Komponente auf, welche Vorverdeckung modelliert. Die Implementation erfolgt entsprechend des

Standards. Nach einer weiteren Unterabstastung um den Faktor 6 wird das in Kapitel 3 vorgestellte interne Rauschen IR addiert. Die Nachverdeckung erfolgt prinzipiell wie im PEAQ FFT Modell (siehe Abschnitt 4.2.2.4), jedoch mit anderen Zeitkonstanten, welche im Standard beschrieben sind. Eine Rücktransformation der Energiewerte in den Zeit- oder Frequenzbereich ist nicht Teil des Standards.

4.4.2 Anpassungen und Erweiterungen

Die im Blockdiagramm in Abbildung 4.5 rot gekennzeichneten Abschnitte (IR und die Rücktransformation) kennzeichnen die Stellen bei der Implementation, die gegenüber den Schritten im Standard angepasst oder hinzugefügt werden.

4.4.2.1 Rücktransformation

Zur ersten Verwendung des Modells wird für die Rücktransformation die des PEAQ FFT Modells angepasst. Die in der Filterbank vorliegenden Signale werden jeweils innerhalb der Bänder superponiert, um entsprechend der Anzahl der Bänder 40 Energiewerte pro Rahmen für die Rücktransformation zu erhalten. Für diese wird die Äquivalentrechteckbandbreitenskala nach [10] verwendet, mit der auch die im Standard vorgegebenen Mittenfrequenzen der Bänder der Filterbank berechnet wurden.

4.4.2.2 Internes Rauschen

Da die erste Implementation keine brauchbaren Ergebnisse liefert, wird die Addition des internen Rauschen herausgenommen. Das Interne Rauschen erhöht, verglichen mit den oben betrachteten Modellen, den im verbesserten Signal der zweistufigen Störgeräuschreduktion verbleibenden Störanteil (verringerte Störgeräuschreduktion). Da die Amplitude des internen Rauschens als Schalldruckpegel gegeben ist und diese im Innenohr mit dem einfallenden Schall addiert wird, ist eine richtige Skalierung des Letzteren unabdingbar. Die übliche Anpassung auf einen entsprechenden Wert von 45 dB Schalldruckpegel verbessert das Ergebnis nicht. Möglicherweise liegt der Grund für die geringe Störgeräuschreduktion an einer anderen Stelle im Modell.

4.5 Bewertung der Modelle im Bezug auf die Anforderungen

Nach der Vorstellung aller behandelten psychoakustischen Modelle, sollen diese kurz im Hinblick auf die genannten Anforderungen (Abschnitt 3.4) bewertet werden.

4.5.1 Repräsentation des Außen- und Mittelohrs und des neuronalen Faktors

Der MPEG1 Layer 3 Standard weist keine vollständige Repräsentation der Filtereigenschaften des Außen- und Mittelohres sowie des neuronalen Faktors auf (Kapitel 3 Abbildung 3.2). Die vom Standard vorgesehene Integration der absoluten Mithörschwelle greift nur bei Situationen, in denen die durch das Sprachsignal erzeugte Mithörschwelle unterhalb der Absoluten liegt. Somit sind die Eigenschaften des Außen- und Mittelohres (3.4

nur für diesen Fall repräsentiert. Der neuronale Faktor als Bestandteil der Gleichungen 3.2 und 3.1 wird gar nicht im Standard repräsentiert.

Hingegen weisen beide Modelle des PEAQ Standards eine Repräsentation des Außen- und Mittelohres auf. Im Zusammenhang mit dem internen Blutrauschen (Abbildung 3.3) wird auch die absolute Mithörschwelle repräsentiert.

4.5.2 Innenohr

4.5.2.1 Abbildung der gehörrichtigen Bänder

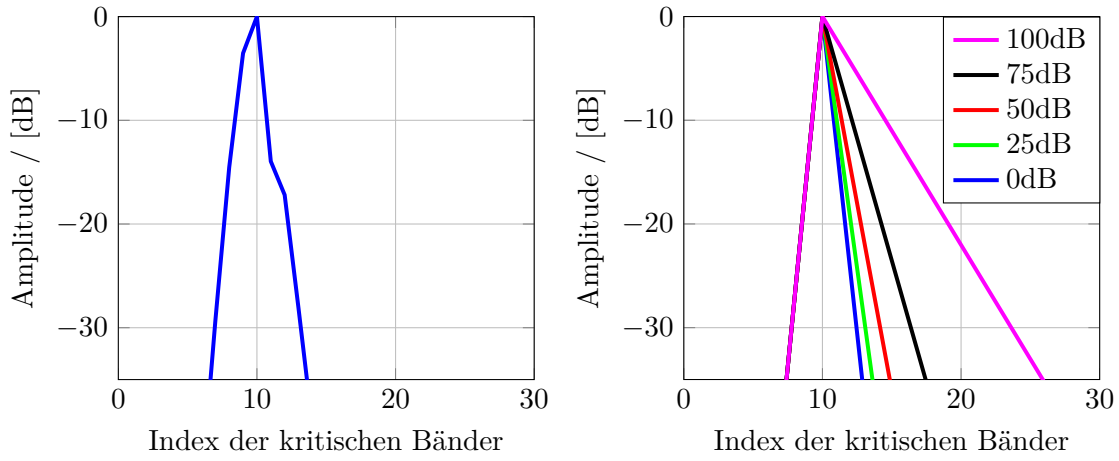
Alle Modelle weisen ähnliche Abbildungen des Signals in gehörrichtige Bänder auf. Die Transformation des FFT Modells des PEAQ Standards nach [28] siehe Abschnitt 3.1.2.2 repräsentiert Untersuchungen zufolge [38] die kritischen Bänder schlechter als die Approximation nach Zwicker und Fastl [14] oder die Äquivalentrechteckbandbreite, welche im Filterbank Modell zum Einsatz kommt. Jedoch soll trotz der schlechteren Repräsentation diese für das FFT PEAQ Modell in der Gesamtleistung besser abschneiden. Im eigenen Test, in denen die Approximation von Schröder nach Gleichung 3.4 durch die von Zwicker 3.5 ausgetauscht wird, lassen sich keine Veränderungen des Klangbildes im verbesserten Signal erkennen. Beim Betrachten der Verläufe der Approximationen in Abbildung 3.5a fällt auf, dass diese erst bei höheren Bändern signifikant voneinander abweichen. Mit der in 2 erwähnten Energieverteilung von Sprache, welche sich fast ausschließlich auf die unteren Bänder erstreckt, lässt sich der nicht wahrnehmbare Unterschied erklären. Möglicherweise wurden bei den Untersuchungen der Unterschiede der Approximationen breitbandigere Audiosignale wie z. B. Musik herangezogen. Die kritischen Bänder beider FFT basierter Modelle weisen in den unteren Bändern oft nur eine zugeordnete Frequenzlinie auf. Die praktische Umsetzung weicht gerade bei diesen wegen der begrenzten Frequenzauflösung von der jeweiligen Approximation ab. Das Filterbank Modell verwendet die Äuivalentrechteckbandbreite wie in Abbildung 3.6 gezeigt. Daher weisen die unteren Bänder schmalere Bandbreiten gegenüber den Approximationen auf. Die Implementierung der Filterbank bietet eine höhere Zeitauflösung für die höheren Bänder gegenüber den FFT Modellen. Die Repräsentation ist wegen der höheren Abtastung im einzelnen Band genauer (Abbildung 4.5). In einem Band werden alle Abtastwerte eines Rahmens entsprechend der Filterbank gewichtet. Bei den FFT Modellen werden hingegen die Frequenzlinien direkt aufsummiert und nur ein Energiewert repräsentiert das jeweilige Band.

Die Bänder sind für alle Modelle nicht überlappend. Den Randeffekten neben der Resonanzstelle auf der Basilarmembran wird durch die Spreizfunktion Rechnung getragen. Das FFT Modell des PEAQ Standards hat gegenüber dem des MPEG1 Layer Standards einen Vorteil. Die Energien, die auf oder in der Nähe der Grenze benachbarter Bänder liegen, werden entsprechend ihrer durch die Frequenzauflösung gegebenen Bandbreite auf die Bänder verteilt. Dies repräsentiert die von Zwicker [43] erwähnte kontinuierlich Einteilung der Basilarmembran besser. Die Anzahl der Bänder liegt bei einer Abtastrate von 48 kHz bei 62 Bändern für den MPEG1L3 Standard und 59 bzw. 109 für das FFT Modell des PEAQ Standards. Letzteres bietet zwei Auflösungen. Die anfangs von Zwicker [43] angeführten 24 Bänder werden auch vom Filterbank Modell (40 Bänder) übertroffen. Da eine Frequenzgruppierung und damit eine Mittelwertbildung der Signale erfolgt, stellt sich die Frage wie weit die Veränderung der Bandbreite der Bänder auf der Tonheitsskala sich

auf das Musical Noise auswirkt.

4.5.2.2 Spreizung und Superposition der Energien

Die Spreizfunktionen der beiden Standards sind in Abbildung 4.6 für das zehnte Band dargestellt. Die logarithmische Amplitude ist gegen den Index der kritischen Bänder aufgetragen. Da die Bänder bzgl. der Tonheitsskala die gleiche Bandbreite dz aufweisen, ist der Index proportional zur Tonheit. Teilabbildung a) zeigt die von der Signalamplitude unabhängige Spreizfunktion des psychoakustischen Modell 2 des MPEG1 Layer 3 Standards. Im Vergleich dazu kann man die Variation der Spreizfunktion des PEAQ Standards in 4.6b für Schalldruckpegel von 0 bis 100 dB sehen. Die Spreizfunktion wird sowohl im FFT Modell als auch in leichter Abwandlung im Filterbank Modell verwendet.



(a) MPEG1 Layer 3 Standard **(b)** PEAQ Standard

Abbildung 4.6: Vergleich der Spreizfunktionen

Beide Spreizfunktionen sind bzgl. der Tonheitsskala in ihrer Form unabhängig. Transformiert in den Frequenzbereich ergibt dies zu hohen Frequenzen hin breitere Spreizfunktionen. Dies resultiert aus der zu hohen Frequenzen hin breiter werdenden Bändern. Als Beispiel sind die Mittenfrequenzen der Bänder in Abbildung 3.7 vorgestellt worden. Das Filterbank Modell weist zusätzlich noch wie oben beschrieben eine Tiefpassfilterung für die Veränderung der Spreizfunktion auf, welche die Latenz des Gehörs, sich auf wechselnde Schalldruckpegel anzupassen, berücksichtigt. Zudem wird das Signal mit Real- und Imaginärteil gespreizt und nicht die Energie. Die Phaseninformation ist bei dem FFT Modellen nicht vorhanden.

Nach [24] superponieren sich die Energien in den einzelnen Bändern nichtlinear 3.14. Diese These wird auch von [6] unterstützt. Nur der PEAQ Standard folgt mit beiden Modellen dieser Einschätzung. Tests haben gezeigt, dass eine lineare Superposition die Klangergebnisse nicht wahrnehmbar verändert. Allerdings ist der verwendete Schalldruckpegel mit 45 dB relativ niedrig. Aus den Maskierungsschwellen Abbildungen 3.9e, welche sich nur im Signal zu Rauschabstand vom Erregungsmuster unterscheiden, zeigen sich erst nichtlineare Phänomene wie Intermodulation bei sehr hohen Schalldruckpegeln. Daher

wirkt sich die nichtlinare Superposition der Energien bei den in dieser Arbeit verwendeten Pegeln kaum auf das Klangergebnis aus. Zusammenfassend kann man sagen, dass das Filterbank Modell zumindest theoretisch gesehen die detailgetreueste Modellierung des Innenohrs darstellt.

4.5.2.3 Temporale Verdeckung

Die temporale Verdeckung ist beim MPEG1 Layer 3 Standard nicht vorgesehen, obwohl sie durch Nachrüsten leichte Verbesserungen mit sich bringt. Die Nachverdeckung des PEAQ Modells stellt eine grobe Approximation dar. Laut Zwicker [14] ist der Verlauf der Mithörschwelle bei der Nachverdeckung nicht exponentiell abfallend, obwohl Zwicker selbst zunächst die Nachverdeckung als exponentiell abfallenden Vorgang modelliert hat 4.6b. Dem FFT Modell des PEAQ Standards wie auch dem MPEG1 Layer 3 Modell fehlt die Modellierung der Vorverdeckung. Diese weist nur das Filterbank Modell auf. Allerdings ist der Effekt der Vorverdeckung aufgrund der kurzen Zeitspanne und der individuellen Empfindung sehr gering. Geübte Hörer nehmen dieses Phänomen kaum bis gar nicht war.

4.5.2.4 Berechnung der Mithörschwellen aus Erregungsmustern

Bezüglich der Berechnung der Maskierungsschwellen zeigt sich nur die Berechnung des MPEG1 Layer 3 Standards als brauchbar. Die im FFT Modell des PEAQ Standards vorgesehen Berechnung resultiert in zu hohen Mithörschwellen aufgrund zu kleiner Signal zu Maskierungsabstände. In der Literatur lässt sich keine theoretische Grundlage für diese finden. Möglicherweise sind die Werte im Hinblick auf das Gesamtmodell angepasst worden. Der eigens eingestellte konstante Signal zu Maskierungsabstand basiert auf der Gleichung (3.15) nach [38]. Das Filterbank Modell des PEAQ sieht keine Berechnung der Mithörschwellen vor.

4.6 Zusammenfassung psychoakustische Modelle

Zusammenfassend kann gesagt werden, dass die Modelle des PEAQ Standards den Anforderungen an ein psychoakustisches Modell eher gerecht werden als die das MP3 Standards. Auch wenn eine Berechnung der Mithörschwellen und eine Tonalitätsberechnung fehlt, ist das Innenohr besonders durch die Amplitudenabhängigkeit der Spreizfunktion und die temporale Verdeckung im PEAQ Standard deutlich besser modelliert. Hinzu kommt das Einbeziehen der spektralen Gewichtung durch das Außen - und Mittelohr sowie des neuronalen Faktors. Das Filterbank Modell besitzt mit den Eigenschaften der Filterbank, der Modellierung der Vorverdeckung, und der Tiefpassgefilterung der Spreizfunktion bei Änderung des Schalldruckpegels die besten Voraussetzungen, das Gehör bestmöglich mathematisch zu beschreiben. Allerdings kann dieser Vorteil bei Anwendung der Rücktransformation des PEAQFFT Modells vom Bereich der kritischen Bänder in den Frequenzbereich nicht ausgenutzt werden. Bei Tests schneidet es schlechter als die FFT Modelle ab. Dies ist auf die unzureichend ausgereifte Rücktransformation und Signalskalierungen zurückzuführen. Das FFT Modell des PEAQ Standards liefert mit allen Kombinationen der im folgenden Kapiteln vorgestellten Filterregeln die besten Ergebnisse bzgl. instrumenteller Maße und informeller Hörtests.

Kapitel 5

Psychoakustische Filterregeln

Im ersten Kapitel wurde in Abschnitt 2.7 die Motivation für die Berechnung der Verdeckungsschwellwerte vorgestellt. Diese sollen in der in Abschnitt 2.7.2 dargestellten zweistufigen Störgeräuschreduktion verwendet werden. Die in Kapitel 4 entwickelten Modelle dienen zur Berechnung dieser Verdeckungsschwellwerte. Dieses Kapitel behandelt die Filterregeln für das Filtergewicht H_{psycho}, die unter anderem die Verdeckungsschwellwerte nutzt. Es haben sich neue Filterregeln ergeben, die anstatt der geschätzten Verdeckungschwellwerte $\hat{R}_{TT}$, die Erregungsmuster E_x und die Lautheit N des geschätzten Sprach- und Störanteils verwenden. Die Lautheit wird aus den Erregungsmustern nach den Gleichungen 3.18 und 3.20 nach [18] berechnet [1].

Die im Kapitel 2 gezeigte Abbildung 2.5, welche die zweistufige Störgeräuschreduktion mit der psychoakustische Stufe(blaue und schwarze Linien) nach Gustafsson [13] beschreibt, kann zum Einbeziehen der neuen Filterregeln entsprechend Abbildung 5.1 abstrahiert werden. Das Wiener Filter und die Störgeräuschschätzung werden zur konventionellen Störgeräuschreduktion zusammengefasst. Für die Verwendung der Erregungsmuster und der Lautheit werden zusätzliche Pfade eingezeichnet (orangene Linien) und bestehende erweitert.

[1] Es können auch andere Modelle zur Berechnung der Lautheit wie z. B. das Modell von Moore und Glasberg [2] verwendet werden, jedoch hat sich bei einem Vergleich [20], das hier verwendete Modell die beste Leistung gezeigt und sich als das Ausgereifteste erwiesen.

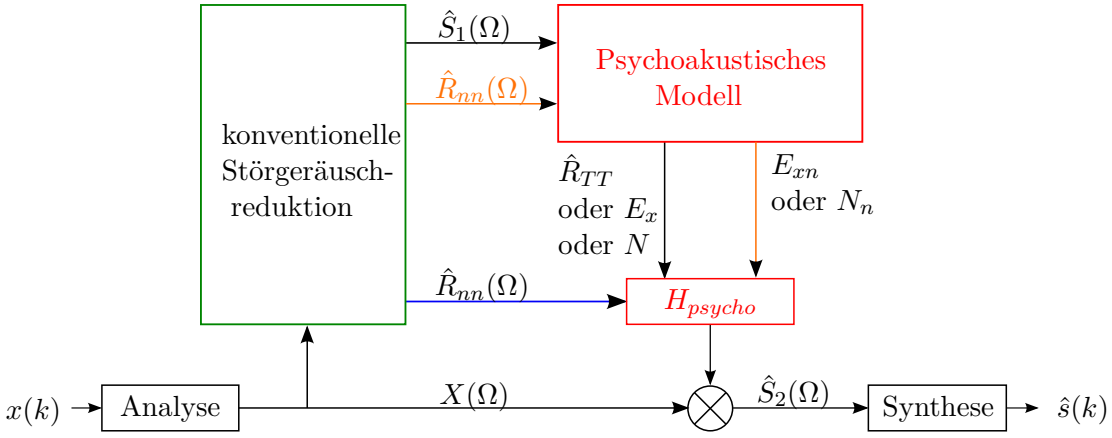


Abbildung 5.1: Verallgemeinerte zweistufige Störgeräuschreduktion

Die Filterregel H_{psycho} der zweiten Stufe (rot gekennzeichnet) wird teilweise zunächst mittels Größen aus dem kritischen Bandbereich berechnet. Die Größen oder das Gewicht selbst (je nach Variante, s.u.) werden anschließend in den Spektralbereich transformiert und sind entsprechend von den Bändern b oder der normierten Frequenz Ω abhängig.

Die bei den in den unteren Abschnitten beschriebenen Filterregeln (für H_{psycho}) verwendeten Pfade im Blockdiagramm (Abbildung 5.1) sind farblich gekennzeichnet. Die Varianten stellen Abwandlungen des Wiener Filters und des von Gustafsson [13] entwickelten H_{IND} nach Gleichung 2.28 dar. Die mit schwarzen Pfeilen versehenen Pfade werden immer ungeachtet der ausgeführten Variante durchlaufen. Der blaue Pfad repräsentiert die Anwendung der ursprünglichen H_{IND} Filterregel. Stattdessen wird für alle neu und weiterentwickelten Filterregeln der orangene Pfad genutzt.

In dieser Arbeit findet für die erste Stufe (grüner Block in Abbildung 5) der Störgeräuschreduktion ausschließlich das Wiener Filter Anwendung. Zunächst soll an die von [13] entworfene Filterregel (Gleichung 5.1) angeknüpft werden.

5.1 HIND Filter

5.1.1 Erregungsbasiertes HIND Filter HINDnEx

Bei der Nutzung der FFT basierten psychoakustischen Modelle (Kapitel 4) erfolgt eine Transformation der Energie des geschätzten Sprachsignals in den Bereich der kritischen Bänder. Die Frequenzgruppierung zu einem Band und die gleichmäßige Verteilung der Energie aus dem kritischen Bandbereich auf die dem Band zugeordneten Frequenzlinien führt zu einer Glättung von Spitzen. Das „Musical Noise“, welches aufgrund von Schätzfehlern des Leistungsspektrums des Störanteils $\hat{R}_{nn}$ hervorgerufen wird, zeigt sich durch charakteristisch vereinzelte Spitzen im Spektrum des verbesserten Signals. Diese Spitzen sind durch Spitzen in $\hat{R}_{nn}$ und darauf basierenden Störabstand zu erklären. Daher liegt es Nahe nicht nur das geschätzte Sprachsignal der ersten Stufe (konventionelle Störgeräuschreduktion) das psychoakustische Modell durchlaufen zu lassen, sondern dies auch für das Leistungsdichtespektrum des geschätzten Störanteils $\hat{R}_{nn}$ vorzunehmen. Allerdings wird für letztere kein Verdeckungsschwellwert berechnet, da nur die Verdeckung des Störanteils

durch den Sprachanteil für die Störgeräuschreduktion interessant ist. Statt dessen werden die Erregungsmustern E_{xgn} von $\hat{R}_{nn}$ berechnet und letzteres damit ersetzt (Gleichung 5.2). Die Berechung nach Gustafsson [13]

$$H_{IND}(\Omega) = \min\left(\sqrt{\frac{\hat{R}_{TT}(\Omega)}{\hat{R}_{nn}(\Omega)}} + \zeta_n, 1\right) \tag{5.1}$$

wird also in folgende überführt:

$$H_{IND,E_x}(\Omega) = \min\left(\sqrt{\frac{\hat{R}_{TT}(\Omega)}{E_{xgn}(\Omega)}} + \zeta_n, 1\right) \tag{5.2}$$

Die untere Schwelle wird entsprechend [13] auf $-20\,\mathrm{dB}$ eingestellt. Wird zur Berechnung der Verdeckungsschwellwerte $\hat{R}_{TT}$ und der Erregungsmuster E_{xgn} das PEAQFFT mit konstantem Signal-Maskierungsabstand SMR verwendet, kann letzterer kleiner gewählt werden, ohne den subjektiven Höreindruck bzgl. des Musical Noise zu verschlechtern. Der kleinere SMR Wert führt zu höheren Verdeckungsschwellwerten. Das Filtergewicht $H_{psycho} = H_{IND,E_x}$ ist dann größer als $H_{IND}(\Omega)$. Die Sprachdämpfung wird reduziert. Für Letzteres sind 14 dB als SMR Wert ausreichend, um gegenüber der konventionellen Störgeräuschreduktion eine signifikante Reduktion von Musical Noise zu vernehmen. Allerdings kann man es bei hoher Abspiellautstärke noch hören. Für das weiter entwickelte Filtergewicht H_{IND,E_x} ist dies bei gleichem SMR Wert weniger hörbar.

5.2 Psychoakustisches Wiener Filter

Das Filtergewicht nach Gustafsson weist eine relativ geringe Störgeräusschreduktion im Vergleich zur konventionellen Störgeräuschreduktion mit Wiener Filter auf. Aus diesem Grunde ergibt sich die Fragestellung, ob die hohe Störgeräuschreduktion des Wiener Filters und das durch Ausnutzung von Psychoakustik stark reduzierte Musical Noise des Filtergewichts nach Gustafsson kombinierbar sind. Daher soll an ein Wiener Filter angelehntes Filtergewicht gesucht werden, welches unter Einbeziehung von Psychoakustik sowohl eine hohe Störgeräuschreduktion wie auch einen niedrigen Grad an Verzerrung (z.B. durch Musical Noise) aufweist.

Das allgemeine Wiener Filter ist mit der Gleichung 5.3 gegeben:

$$H_{Wiener}(D) = \frac{\hat{\eta}_2(D)}{1 + \hat{\eta}_2(D)} \tag{5.3}$$

Der apriori Störabstand $\hat{\eta}_2$ kann sowohl im Frequenzbereich als auch im Bereich der kritischen Bänder berechnet werden. Die Variable D ist ein Platzhalter für die Variablen b (kritische Bänder) oder die Ω (Frequenzbereich). Zu beachten ist, dass es sich hier um ein weiteres Wiener Filter (zweite Stufe H_{psycho}) als Nachschaltung der konventionellen Störgeräuschreduktion (erstes Wiener Filter) handelt. Der Störabstand $\hat{\eta}_2$ wird bei jeder hier vorgestellten Variante nicht wie bei der ersten Stufe mit dem „Decision Directed“ Ansatz berechnet, da dies zu wahrnehmbar schlechteren Ergebnissen im Klang führt. Dies

lässt sich damit erklären, dass Größen, welche Verdeckungsphänomene enthalten, wie Erregungsmuster, Verdeckungsschwellwert und Lautheit oder darauf basierende Größen wie der Störabstand nicht zeiltich oder spektral zusätzlich „gespreizt“ bzw. „verschmiert“ werden sollten. Die Repräsentation des Gehörs ginge verloren.

Bei den ersten Versuchen fällt bei erregungsbasierten Berechnungen des Störabstands auf, dass das Filtergewicht für kleine Störabstände zu groß ist und Verzerrungen zu hören sind. Daher wird dies durch einen zusätzlichen Parameter k_w eingestellt:

$$H_{2,Wiener} = \frac{\hat{\eta} - k_w}{1 + \hat{\eta} - k_w} \tag{5.4}$$

Das Filtergewicht $H_{2,Wiener}$ wird dabei auf den Bereich 0...1 begrenzt, auch um negative Gewichte zu verhindern. Die Modifikation ist gerade für die sehr kleinen Störabstande gedacht, da die Störgeräuschschätzung der ersten Stufe bei diesen die größten Abweichungen zum tatsächlichen Störgeräuschanteil aufweist. Der Schätzfehler des Leistungsdichtespektrums führt auch beim „Decision Directed“ Ansatz nach Gleichung 2.20 zu einem Fehler des daraus ermittelten Leistungsdichtespektrums des geschätzten Sprachanteils. Da nun beide fehlerbehafteten Größen zur Bestimmung des Störabstands der ersten Stufe verwendet werden, pflanzt sich der Fehler beim Störabstand fort. Der in der ersten Stufe „Decision Directed“ verwendete Ansatz wirkt diesem Phänomen durch Glättung etwas entgegen. Da der Parameter relativ klein gewählt wird und nur bei kleinen Störabständen greifen soll, soll versucht werden die Regel durch Approximation zu vereinfachen. Dies führt auf die ursprüngliche Wiener Filterregel nach Gleichung 5.3. Allerdings wird zusätzlich ein Schwellwert für den Störabstand eingeführt. Fällt letzterer unterhalb dieses Schwellwerts, wird er zu 0 bzw. 0,001...0,1 (entsprechend $-30..10\,\mathrm{dB}$) gesetzt.

Zunächst sollen die verschiedenen Varianten, welche in den folgenden Abschnitten vorgestellt werden, anhand der Abbildung 5.2 kategorisiert werden. Die auf dem geschätzten Sprachsignal basierende wird mit A, die auf dem geschätzten Störanteil basierende Größe mit B bezeichnet. Die Größen A und B können je nach Variante die Erregungsmuster oder die Lautheiten des geschätzten Sprach- und Störanteils annehmen. In der Abbildung sind drei mögliche Berechnungspfade dargestellt.

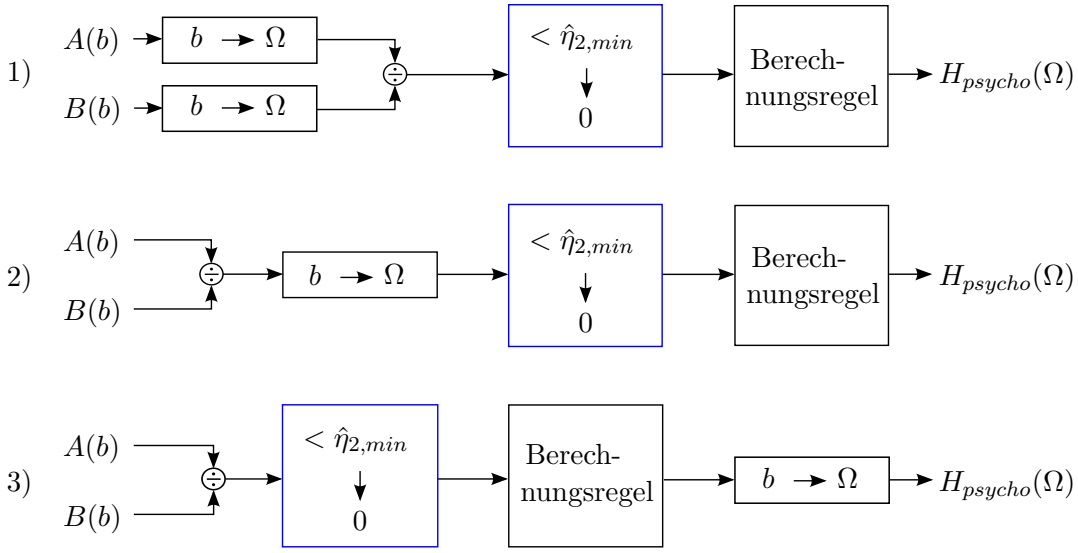


Abbildung 5.2: Varianten zur Berechnung der psychoakustischen Wiener Filterregeln

Da die Transformation vom Frequenzbereich in den Bereich kritischer Bänder und umgekehrt etwas durch die Segmentierung in Rahmen fehlerbehaftet ist und teilweise Artefakte generiert, werden die Berechnungen für den Störabstand $\hat{\eta}_2$ und das Filtergewicht H_{psycho} in den beiden Bereichen untersucht. Alternativ könnte man auch die Fensterung der Rahmen ändern, dies verzerrt allerdings die im Bereich der kritischen Bänder berechneten Energiewerte. sodass von dieser Möglichkeit kein Gebrauch gemacht wird. Besonders interessant sind mögliche unterschiedliche Leistungen der Varianten und Filterregeln in der Störgeräuschreduktion bzgl. weißem Rauschen oder des Bohrhammers als Störanteil. Gesucht wird eine Variante, die ähnlich gute Ergebnisse für beide Fälle liefert.

Allgemein folgen die drei Pfade in Abbildung 5.2 den folgenden Schritten, wenn auch in unterschiedlicher Reihenfolge:

- Rücktransformation der entsprechenden Größe(n) vom Bereich der kritischen Bänder in den Frequenzbereich. Da das PEAQFFT Modell mit konstantem SMR die besten Ergebnisse liefert, wird dieses als einziges Modell mit den in diesen Kapiteln vorgestellten Filterregeln kombiniert. Daher erfolgt die Transformation entsprechend der Rücktransformation nach Schröder [28] (Gleichung 3.7).
- optionaler Vergleich des Störabstands $\hat{\eta}_2$ mit einem Schwellwert $\hat{\eta}_{2,min}$, der ein Minimum des Störabstandes festlegt. Dies ist in dem Block symbolisch mit 0 gekennzeichnet. Es ist anzumerken, dass $\hat{\eta}_{2,min}$ verschiedene Werte je nach verwendeter Größe (Lautheit oder Erregungsmuster) sowie in Abhängigkeit des Bereichs annehmen kann.
- Berechnung des Filtergewichts H_{psycho} nach der Wiener Filter Regel.

Die Reihenfolge der Schritte bestimmt, welche Größe(n) durch die Rücktransformation direkt geglättet werden. Im ersten Pfad werden die Größen A und B durch Rücktransformation geglättet. Diese Größen repräsentieren Erregungsmuster oder Lautheiten des

Sprach-und Störanteils. Hingegen erfährt der Störabstand bei Verwendung des zweiten Pfades eine Glättung, in der dritten Variante das Filtergewicht $H_{psycho}(b)$. Interessant wäre auch die Gewichtung selbst im Bereich der kritischen Bänder vorzunehmen, d.h. das verrauschte Signal $X(\Omega)$ in $X(b)$ zu überführen und mit dem Filtergewicht $H_{psycho}(b)$ zu multiplizieren und anschließend erst die Rücktransformation vorzunehmen. Dies käme der Verarbeitung von verrauschten Signalen im menschlichen Gehör am nächsten, wird allerdings nicht untersucht.

5.3 Erregungsbasiertes Wiener Filter

Für den erregungsbasierten Wiener Filter werden alle Varianten aus Abbildung 5.2 getestet.

5.3.1 Berechnung des Störabstands im Frequenzbereich (WienerExSNRw Pfad 1)

Die Erregungsmuster des Sprach- und Störsignals $E_{xgs}(b)$ und $E_{xgn}(b)$ werden in den Frequenzbereich transformiert. Der Störabstand $\hat{\eta}_2$ berechnet sich nach:

$$\hat{\eta}_2(\Omega) = \frac{E_{xgs}(\omega)}{E_{xgn}(\Omega)} \tag{5.5}$$

Optional kann zur weiteren Reduktion des Musical Noise der Schwellwert des Störabstands $\hat{\eta}_2$ auf den Wert 0.4 gesetzt werden. Das Filtergewicht $H_{psycho}(\Omega) = H_{2,Wiener,E_x}(\Omega)$ ergibt sich wie folgt:

$$H_{2,Wiener,E_x}(\Omega) = \frac{\hat{\eta}_{2,N}(\Omega)}{1 + \hat{\eta}_{2,N}(\Omega)} \tag{5.6}$$

5.3.2 Berechnung des Störabstands im Bereich kritischer Bänder (WienerExSNRb Pfad 2)

Die Erregungsmuster werden nicht transformiert. Stattdessen wird der Störabstand direkt bestimmt,

$$\hat{\eta}_2(b) = \frac{E_{xgs}(b)}{E_{xgn}(b)} \tag{5.7}$$

Auch hier kann mit dem Schwellert das restliche Musical Noise reduziert werden ($\hat{\eta}_2 = 0.3$). Anschließend wird der Störabstand in den Frequenzbereich transformiert und das Filtergewicht bestimmt.

$$H_{2,Wiener,E_x}(\Omega) = \frac{\hat{\eta}_{2,N}(\Omega)}{1 + \hat{\eta}_{2,N}(\Omega)} \tag{5.8}$$

5.3.3 Berechnung der Filterregel im Bereich kritischer Bänder (WienerbExSNRb Pfad 3)

In der dritten Variante wird der Störabstand wie in Gleichung 5.7 bestimmt. Der Schwellert wird auf $\hat{\eta}_2 = 0.3$ eingestellt. Das Filtergewicht wird hier im Bereich der kritischen Bänder

berechnet:

$$H_{2,Wiener,E_x}(b) = \frac{\hat{\eta}_{2,N}(b)}{1+\hat{\eta}_{2,N}(b)} \tag{5.9}$$

5.3.4 Verbessertes erregungsbasiertes Wiener Filter (WienerExSNRb2)

Die Idee, Artefakte aufgrund der Transformationen mittels der Verschiebung von Teilen der Berechnung des Filtergewichts in den Bereich der kritischen Bänder zu reduzieren, hat nur einen geringen Einfluss auf die Ergebnisse. Die Verwendung des Schwellwerts führt teilweise zu Nachhalleffekten und Verzerrungen, welche auf sprunghaftes Abfallen des Störabstands für Werte unterhalb des Schwellwerts zurückzuführen sind. Zudem ist wäre eine stärkere Störgeräuschreduktion wünschenswert. Das Filtergewicht $H_{psycho} = H_{2,Wiener,E_x}$ soll daher bei geringen Störabständen noch kleiner und bei hohen Störabständen, falls möglich, größer oder zumindest nicht verringert werden. Dies führt bei den Tests zu folgendem allgemeinen Filtergewicht H_{2,W,E_x}

$$H_{2,W,E_x} = \left(\frac{\hat{\eta}_{2,N}^{e_1}}{1+\hat{\eta}_{2,N}^{e_2}}\right)^{e_3} \tag{5.10}$$

Eine erhöhte Störgeräuschreduktion ergibt sich für die Werte der Exponenten $e_1 = 1$, $e_2 = 1$, $e_3 = 1...2$. Die Gleichung vereinfacht sich zu:

$$H_{2,W,E_x} = \left(\frac{\hat{\eta}_{2,N}}{1+\hat{\eta}_{2,N}}\right)^{e_3} \tag{5.11}$$

Besonders stark profitiert der Höreindruck des Reststöranteils bei transientem Störanteil (Bohrhammer) gegenüber der ursprünglichen Variante WienerExSNRb . Allerdings führen Werte von $e_3 > 1{,}7$ zu wahrnehmbaren Verzerrungen. Bei dieser Variante wird auch kein Schwellwert herangezogen, um niedrige Werte des Störabstands weiter zu dämpfen. Mit der Potenzierung des Filtergewichts reduzieren sich die Artefakte, welche durch Anwenden der Schwellwerte bei den ersten drei erregungsbasierten Regeln verursacht werden. Die Störgeräuschreduktion nimmt merklich zu. Insgesamt ist die letzte Variante sowohl für den Fall des transienten als auch für den des stationären Störanteils die beste Variante unter den erregegungsbasierten Wiener Filtern.

5.4 Lautheitsbasiertes Wiener Filter

Das lautheitsbasierte Wiener Filter soll der empfundenen Lautstärke des Sprach- und Störanteils bzw. deren Verhältnis zueinander Rechnung tragen. Ein Verhältnis der Schalldruckpegel beider Anteile berücksichtigt nicht die Psychoakustik des Gehörs, welche einen Anstieg des Schalldruckpegels nicht linear mit diesem wahrnimmt. Die empfundene Lautstärke (auch als Lautheit bezeichnet) wird in Abschnitt 3.3.2 beschrieben.

Für das lautheitsbasierte Wiener Filter werden nur die beiden Varianten 1) und 2) in Abbildung 5.2 untersucht.

Die Lautheit wird nach dem Modell von Zwicker [43] [2] unter Verwendung der durch das psychoakustische Modell PEAQFFT bestimmten Erregungsmuster berechnet. Der Störabstand $\hat{\eta}_{2,N}(b)$ setzt die Lautheiten des Sprach- und Störanteils ins Verhältnis (Gleichung 5.14). Die zweistufige Störgeräuschreduktion nach Abbildung 5.1 durchläuft daher die schwarzen und orangenen Pfade.

5.4.1 Berechnung des Störabstands im Frequenzbereich (WienerNSNRw Pfad 1)

Zunächst werden die Lautheiten $N_s(b)$ und $N_n(b)$ des geschätzten Sprach- und Störanteils entsprechend der Variante 1) in der Abbildung 5.1 in den Frequenzbereich transformiert. Anschließend erfolgt die Berechnung des Störabstands $\hat{\eta}_{2,N}(\Omega)$:

$$\hat{\eta}_{2,N}(\Omega) = \frac{N_s(\Omega)}{N_n(\Omega)} \tag{5.12}$$

Der Schwellwert $\hat{\eta}_{2,N,min}$ für den alle darunter liegenden Werte des Störabstands auf 0.001 gesetzt werden, liegt bei 0.6 ($-2.22\,\mathrm{dB}$). Zuletzt wird die Filterregel der zweiten Stufe $H_{psycho}(\Omega) = H_{2,Wiener,N}(\Omega)$ bestimmt.

$$H_{2,Wiener,N}(\Omega) = \frac{\hat{\eta}_{2,N}(\Omega)}{1 + \hat{\eta}_{2,N}(\Omega)} \tag{5.13}$$

Das Ergebnis weist eine zu geringe Störgeräuschreduktion auf.

5.4.2 Berechnung des Störabstands im Bereich krit. Bänder (WienerbNSNRb Pfad 2)

$$\hat{\eta}_{2,N}(b) = \frac{N_s(b)}{N_n(b)} \tag{5.14}$$

Der Schwellwert für den Störabstand wird auf $\hat{\eta}_{2,N}(b) = 0.5$ ($-3\,\mathrm{dB}$) eingestellt. Alle darunter liegenden Werte des Störabstands werden auf 0.001 ($-30\,\mathrm{dB}$) gesetzt. Anschließend wird der Störabstand in den Frequenzbereich transformiert und das Filtergewicht ermittelt:

$$H_{2,Wiener,N}(\Omega) = \frac{\hat{\eta}_{2,N}(\Omega)}{1 + \hat{\eta}_{2,N}(\Omega)} \tag{5.15}$$

5.4.3 Verbessertes lautheitsbasiertes Filter (WienerNSNRb2 Pfad 2)

Nur Variante WienerSNRNw liefert für stationäre Rauschanteile akzeptable Ergebnisse. Daher wird die ursprüngliche Anpassung des Wiener Filters (Gleichung 5.16) etwas flexibler gestaltet. Die Berechnung wird ausschließlich entsprechend des Pfad 2) in Abbildung 5.2 durchgeführt, da durch eine Glättung des Störabstands (durch die Transformation) ein besseres Ergebnis für den Fall des transienten Störanteils erwartet wird.

$$H_{2,W,N}(\Omega) = \frac{\hat{\eta}(\Omega) - k_{w1}}{1 + \hat{\eta}(\Omega) - k_{w2}} \tag{5.16}$$

[2] entspricht dem ISO Standard 226 [18]

Die Veränderung des Zählers und Nenners sind getrennt über die Parameter k_{w1} und k_{w2} durchführbar. Tests führen sowohl für transiente Störgeräusche (Bohrhammer) wie auch für stationäre Störgeräusche(weisses Rauschen) auf $k_{w1} = 1.5$ und $k_{w2} = 1$. Damit lässt sich die Filterregel $H_{2,W}$ vereinfachen:

$$H_{2,W,N}(\Omega) = \frac{\hat{\eta}(\Omega) - 1.5}{\hat{\eta}(\Omega)} \tag{5.17}$$

Das Filtern niedriger Werte für den Störabstand über die Schwellwertregelung (blau umrahmter Block in Abbildung 5.2) entfällt. Die Filtergewichte $H_{2,W,N}(\Omega)$ werden auf den Bereich 0 ... 1 begrenzt, um negative Gewichtungen und Verstärkungen zu verhindern. Das Ergebnis ist für den stationären Störanteil noch besser, es weist weniger Artefakte wie z. B. „Knackser“ auf. Gegenüber den oben vorgestellten lautheitsbasierten Filterregeln reduziert sich das „Musical Noise“ bei transientem Störanteil signifikant.

5.5 Zusammenfassung psychoakustische Filterregeln

Alle vorgestellten Filterregeln nutzen Ansätze zur Repräsentation der Psychoakustik. Die Filterregeln lassen sich in vier Kategorien gruppieren:

1. Die ursprüngliche Filterregel setzt das Leistungsdichtespektrum des Verdeckungsschwellwerts Sprachanteils mit dem Leistungsdichtespektrum des geschätzten Störanteils ins Verhältnis.

2. Das erregungsbsierte HIND Filter erweitert die psychoakustische Repräsentation, durch Austauschen des Leitungsdichtespektrum des Störanteils mit dessen Erregungsmuster. Im Vergleich zu 1. werden die Größen des Verhältnisses (Sprach- zu Störanteil) rein psychoakustisch berechnet.

3. Das erregungsbasierte Wiener Filter nutzt das Verhältnis der Erregungsmuster des geschätzten Sprach- und Störanteils zur Berechnung des Störabstands. Dieser bezieht damit die psychoakustische Repräsentation von Schall im Innenohr in eine konventionelle Filterregel mit ein.

4. Das lautheitsbasierte Wiener Filter versucht das Phänomen der empfundene Lautstärke abzubilden. Die Lautheiten des geschätzten Sprach- und Störanteils werden auf Basis der jeweiligen Erregungsmuster berechnet und ins Verhältnis gesetzt.

Kapitel 6

Ergebnisse

In diesem Kapitel sollen ausgewählte Kombinationen der psychoakustischen Modelle (Kapitel 4) und der Filterregeln (Kapitel 5) anhand der in Abschnitt 2.4 vorgestellten instrumentellen Maße objektiv bewertet werden. Die Motivation zur Anwendung einer psychoakustischen Störgeräuschreduktion Abschnitt 2.7) besteht aus folgenden Aspekten:

- Reduktion der Sprachdämpfung durch Ausnutzung von Psychoakustik zur Erhöhung der Sprachverständlichkeit
- Minimierung von Artefakten und Sprachverzerrungen mit dem Fokus auf „Musical Noise“
- wahrgenommene Störgeräuschreduktion dabei möglichst auf dem Niveau der konventionellen Störgeräuschreduktion halten.

Unter diesen Gesichtspunkten sollen die Verfahren miteinander verglichen werden. Die konventionelle Störgeräuschreduktion mit einem Wiener Filter wird als Referenz für alle Varianten der zweistufigen Störgeräuschreduktion genutzt.

6.1 Vorgehen

Die Bewertung lässt sich in zwei Teile gliedern. Im ersten Teil werden die FFT basierten psychoakustischen Modelle gegenüber gestellt. Die konventionelle Störgeräuschreduktion mit dem Wiener Filter dient als Referenz für die zweistufige Störgeräuschreduktion. Die psychoakustischen Modelle werden in diesem Teil mit der von Gustafsson [13] entwickelten HIND Filterregel (Gleichung 5.1) kombiniert. Im zweiten Teil werden die vorgestellten Filterregeln (Kapitel 5) miteinander verglichen. Dazu wird das beste psychoakustische Modell aus dem ersten Teil für die Berechnung der Filterregeln innerhalb der zweistufigen Störgeräuschreduktion verwendet.

6.2 Benchmark

Die Leistung der zweistufigen Störgeräuschreduktion (psychoakustische Modelle und Filterregeln) wird durch Durchführung eines Benchmarks gemessen. Die instrumentellen Ma-

ße werden unter Variation folgender Größen bestimmt:

- Der Störabstand wird von −10 dB bis 15 dB in 5 dB Schritten erhöht.
- Jedes Störgeräuschreduktionssystem wird für jeweils zwei weibliche und männliche Sprecher durchlaufen. Die Dateien sind aus der institutseigenen Audio Datenbank [1] entnommen.
- Als Störsignal werden zwei Extremfälle gewählt, um die Grenzen transienter und stationärer Charakteristik einzubeziehen. Das transiente Störsignal ist durch das Bohrhammer Signal aus der Institutsdatenbank [2] repräsentiert. Als stationäres Störsignal wird weißes Rauschen mit einer Abtastfrequenz von 48 kHz selbst generiert.

Das verrauschte Signal wird mit einem Tiefpassfilter mit einer Grenzfrequenz von 16 kHz gefiltert. Dies ist erforderlich, um zwischen der konventionellen Störgeräuschreduktion und der der zweistufigen Störgeräuschreduktion Vergleichbarkeit herzustellen. Letztere enthält die psychoakustischen Modelle, welche nur Signale mit Bandbreiten zwischen 16 und 18 kHz verarbeiten. Auch die Abtastfrequenz der den psychoakustischen Modellen zugeführten Spektren ist durch deren Standards auf 48 kHz festgelegt.

6.3 Bewertung der psychoakustischen Modelle

6.3.1 Auswahl psychoakustischer Modelle

Zunächst wird die konventionelle Störgeräuschreduktion mit dem Wiener Filter mit der zweistufigen Störgeräuschreduktion unter Nutzung des HIND Filtergewichts nach Gustafsson [13] Gleichung miteinander verglichen. Dabei kommen die FFT basierten psychoakustischen Modelle 1 bis 3 zum Einsatz:

1. Das MP3G Modell stellt die psychoakustische Referenz dar, an der die Leistungsverbesserungen nachfolgend untersuchter Modelle aufgezeigt werden soll. Es basiert auf dem von Gustafsson [13] verwendeten MPEG1 Layer 3 Model (Abschnitt 4.1) mit kleineren Anpassungen zur Einbettung in das verwendete Störgeräuschreduktionssystem.
2. Das erweiterte an das MP3G anknüpfende Modell MP3GADV verwendet das Minimum der Verdeckungsschwellwerte wie in Abschnitt 4.2.2.3 beschrieben. Zusätzlich ist die temporale Verdeckung des nachfolgenden Abschnitts 4.2.2.4 Teil des Modells.
3. Das psychoakustische Modell des PEAQ Standards wird durch das PEAQFFT Modell repräsentiert, welches einen konstanten SMR wie in Abschnitt 4.3.2.5 beschrieben verwendet. Dieser zur Berechnung der Verdeckungsschwellwerte aus den Erregungsmustern verwendete SMR wird auf 14 dB eingestellt.

Die letzten beiden Modelle stellen jeweils die besten Erweiterungen des jeweiligen Standards dar.

[1] Pfad Sprecher: '/share/sounds/databases/TSPspeech/48k/'

[2] Pfad Störgeräusch:'/share/sounds/databases/ETSI_EG_202_396_1_Background_Noise/Binaural_Signals/Work_Noise_Jackhammer_binaural (0.00-20.00 s).wav)'

6.3.2 Sprachverzerrung (cepstrale Distanz)

Zur Messung der Sprachverzerrung wird die cepstrale Distanz 2.4 gewählt, da diese die unteren Frequenzen des Sprachsignals stärker gewichtet. Letztere weisen bei der Energieverteilung über dem Frequenzspektrum für die unteren Frequenzen bis ca. 4 kHz den Großteil ihrer Energie auf. In der Abbildung 6.1 ist die cepstrale Distanz zwischen originalen Sprachsignal s und dessen gefilterten Signal $\tilde{s}$ für die konventionelle Störgeräuschreduktion und für die zweistufige Störgerauschreduktion unter Verwendung der psychoakustischen Modelle aufgetragen.

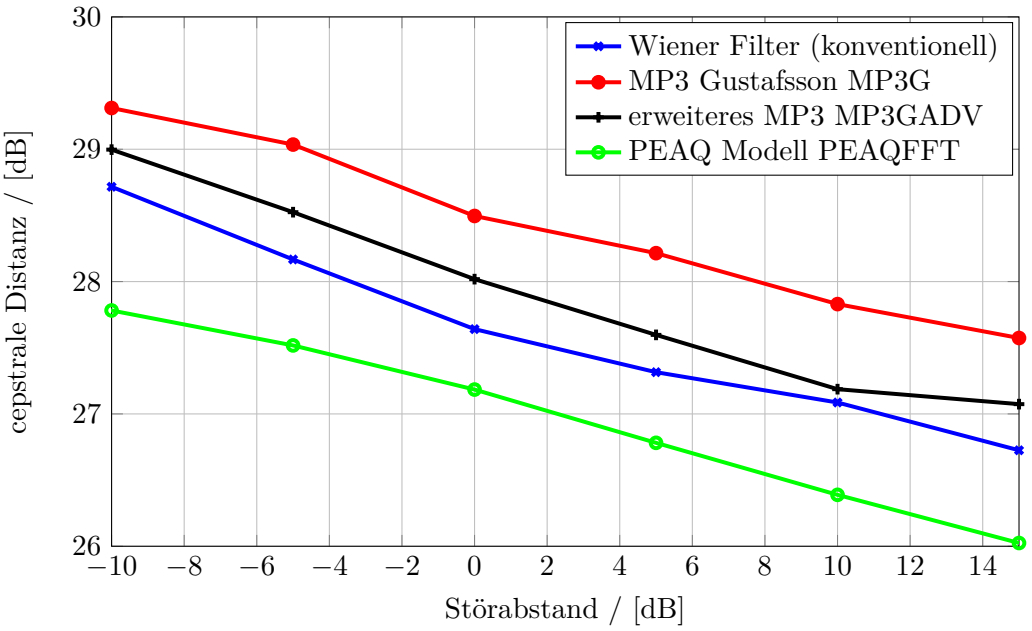


Abbildung 6.1: Vergleich der Sprachverzerrung der konventionellen und der zweistufigen psychoakustisch basierten Störgeräuschreduktion (MP3G, MP3GADV, PEAQFFT)

Die Verläufe der cepstralen Distanz mit zunehmenden Störabstand sind für alle Varianten näherungsweise parallel. Während der cepstrae Distanz des Wiener Filter im unteren Mittelfeld liegt, weist die zweistufige Störgeräuschreduktion mit dem Modell MPG3G nach Gustafsson die höchsten Werte auf. Im niedrigen Frequenzbereich ist bei diesem Modell die Verzerrung der Sprache am stärksten. Der Verlauf des erweiterten Modells MP3GADV liegt 0,5 bis 1 dB unterhalb von dem des Modells MP3G. Dies ist auf die zweifache Berechnung der Verdeckungsschwellwerte und der zusätzlichen temporalen Verdeckung zurückzuführen, welches die Genauigkeit der aus Schätzgrößen bestimmten Verdeckungsschwellwerte erhöht. Dies resultiert in höheren Filtergewichten, an den Stellen im Signal, wo es durch Verdeckung möglich ist, einen höheren Rauschanteil gegenüber einer Filterung mit konventioneller Störgeräuschreduktion zuzulassen. Das Ziel mittels einer zweistufigen Störgeräuschreduktion die Sprachverzerrung zu minimieren und damit die cepstrale Distanz, wird von den MP3 Modellen nicht erreicht. Hingegen erreicht das PEAQFFT Modell mit einem Abstand von 1 bis 2 dB eine Verbesserung gegenüber dem Wiener Filter. Die Redu-

zierung der Sprachverzerrungen ist über einen informellen Hörtest gut wahrnehmbar. Zu erklären ist dies mit der deutlich genaueren Repräsentation psychoakustischer Phänomene (Abschnitt 4.3) durch das PEAQFFT Modell gegenüber den MP3 Modellen.

6.3.3 Störgeräuschreduktion Differenz zwischen segmenteller Rausch- und Sprachdämpfung

In der oberen Abbildung 6.1 hat sich gezeigt, dass eine Verringerung der Sprachverzerrung durch Verwendung einer zweistufigen psychoakustisch basierten Störgeräuschreduktion ermöglicht wird. Im Allgemeinen geht mit einer geringeren Sprachverzerrung eine geringere Sprachdämpfung einher. Im Abschnitt 2.7 ist in der Abbildung 2.4 das Filtergewicht einer allgemein betrachteten Störgeräuschreduktion in Abhängigkeit der Sprach- und Störgeräuschdämpfung aufgetragen. Wie man sieht, lässt sich grundsätzlich die Sprachdämpfung verringern, indem die Störgeräuschdämpfung gesenkt wird. Es stellt sich nun die Frage, ob die Reduktion der Sprachverzerrung mittels Psychoakustik zu einem schlechteren Rausch- zu Sprachdämpfungsabstand $SegDA$ (Gleichung 2.10), bzw. linear betrachtet zu einem schlechteren Verhältnis von Rausch zu Sprachdämpfung, führt. Die verringerte Störgeräuschreduktion (Störgeräuschdämpfung) ist unter Anwendung von Psychoakustik für Zeitbereiche hoher Sprachsignalamplitude entsprechend den Erläuterungen in Abschnitt 2.7 beabsichtigt. Zudem ist die „wahrgenommeneSStörgeräuschreduktion interessant. Wird die Sprachdämpfung verringert und das Sprachsignal aufgrund der höheren mittleren Amplitude besser verständlich, so wird bei gleichem Niveau des Störgeräuschanteils, dieser wegen der in Abschnitt 3.2 beschriebenen Verdeckungseffekte als weniger stark empfunden. Daher eignet sich ein Vergleich der Stördämpfungen von konventioneller und psychoakustisch basierter Störgeräuschreduktion nur bedingt. Aus diesem Grund wird statt der Stördämpfung die Differenz der Dämpfungen in Abhängigkeit zum Störabstand in Abbildung 6.2 betrachtet.

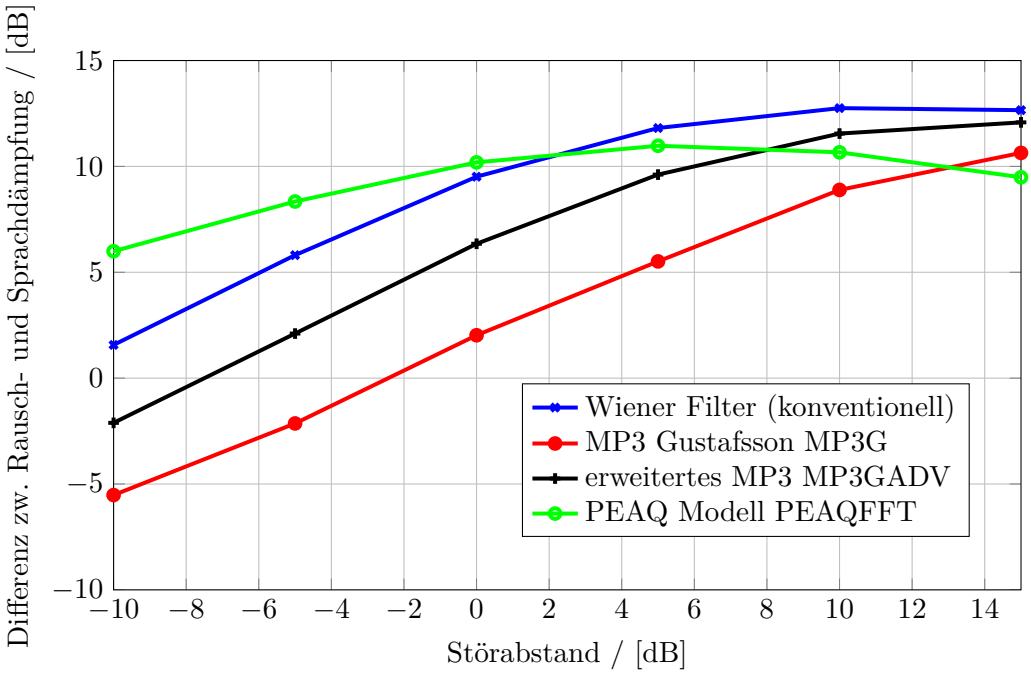


Abbildung 6.2: Differenz zwischen Rausch- und Sprachdämpfung in Abhängigkeit des Störabstands, Vergleich der psychoakustischen Modelle der zweistufigen mit der konventionellen Störgeräuschreduktion

Erstrebenswert sind hohe Werte für die Differenz, da dies einer relativ hohen Störgeräuschdämpfung verglichen mit der Sprachdämpfung entspricht. Für niedrige Störabstände bis ca. 4 dB verlaufen die Differenzen des Wiener Filters und der zweistufigen Störgeräuschreduktion mit den Modellen MP3G und MP3GADV nahezu parallel. Dabei weist das Modell MP3G zu der konventionellen Störgeräuschreduktion mittels Wiener Filter eine um bis zu 7 dB geringere Differenz der Dämpfungen auf. Die Erweiterung MP3GADV halbiert den Abstand zum Wiener Filter gegenüber dem Modell MP3G. Das PEAQFFT Modell zeigt bis zu einem Störabstand von 2 dB sogar höhere Werte als mit der konventionellen Störgeräuschreduktion erzielt wird. Der Unterschied zu letzterem ist bei niedrigen Störabständen besonders groß (bis zu 5 dB), da hier die Verdeckung voll ausgenutzt werden kann. Die Ergebnisse korrelieren mit den Verläufen der cepstralen Distanz in Abbildung 6.1. Je besser die psychoakustische Repräsentation, desto höher das Dämpfungsverhältnis. Während sich Dämpfungsverhältnisse der konventionellen und der zweistufigen Störgeräuschreduktion mit den MP3 Modellen für hohe Störabstände (8 dB) annähern, sinkt das des zweistufigen Störgeräuschreduktion mit dem PEAQFFT Modell ab. Für hohe Störabstände eignet sich der konstant gewählte Signal-Maskierungsabstand des PEAQFFT Modells insofern nicht, als das zwar die Psychoakustik repräsentiert ist, allerdings ist eine Verdeckung des Störanteils durch den Sprachanteil in nur noch geringem Maße nötig. Bei der betrachteten zweistufigen Störgeräuschreduktion kommt das H_{IND} Filtergewicht zum Einsatz, welches das Leistungsdichtespektrum des Verdeckungsschwellwert des geschätzten Sprachsignals mit dem Leistungsdichtespektrum des Störanteils ins Verhältnis setzt. Durch den konstanten Signal Maskierungsabstand von 14 dB ist der Verdeckungsschwellwert und

damit das Filtergewicht nach oben begrenzt. Nimmt der Störabstand den Wert des Signal Maskierungsabstandes an, wird der Verdeckungseffekt nicht mehr voll ausgenutzt. Die Maskierschwelle könnte mehr verdecken als den vorhandenen Störanteil. Gleichzeitig kann aber der Zähler des HIND Filters (Gleichung 5.1) nicht über die Wurzel des Leistungsdichtespektrums des geschätzten Sprachanteils abzüglich des Signal Maskierungsabstands von 14 dB steigen. Für die MP3 Modelle hingegen ist letzterer tonalitätsabhängig und kann deutlich niedrigere Werte bis 5 dB annehmen. Dadurch ergeben sich für letztere höhere Filtergewichte und eine damit einhergehende niedrigere Sprachdämpfung. Da die Störgeräuschdämpfung für höhere Störabstände abnimmt, fällt das Verhältnis der Dämpfungen der MP3 Modelle bei diesen höher aus.

Das PEAQFFT Modell eignet sich zum Einsatz in einer zweistufigen Störgeräuschreduktion für eher geringe Störabstände bis 5 dB.

6.4 Bewertung der psychoakustischen Filterregeln

6.4.1 Auswahl psychoakustischer Filterregeln

In diesem Abschnitt sollen die psychoakustischen Filterregeln unter Verwendung des psychoakustischen Modells PEAQFFT mit einem konstanten SMR von -14 dB miteinander verglichen werden, da dies die geringste Sprachverzerrung und eine gutes Verhältnis von Rausch- zu Sprachdämpfung aufweist (Abschnitt 6.3). Zusammen bildet das psychoakustische Modell mit der jeweiligen Filterregel die in Abbildung 5.1 gezeigte zweite Stufe der zweistufigen Störgeräuschreduktion.

Es werden die Filterregeln aus folgenden Kategorien 2. bis 4. (Abschnitt 5.5) ausgewählt. Kategorie 1. und 2. liefern bei der Bewertung durch die instrumentellen Maße marginale Unterschiede. Somit sind prinzipiell alle Kategorien abgedeckt. Es werden drei Filterregeln dem Wiener Filter der konventionellen Störgeräuschreduktion gegenübergestellt:

- der erregungsbasierte HIND Filter, welcher den Verdeckungsschwellwert des geschätzten Sprachsignals mit dem Erregungsmuster des geschätzten Leistungsdichtespektrums des Störsignals ins Verhältnis setzt.
- aus der dritten Kategorie wird das verbesserte erregungsbasierte Wiener Filter WienerExSNRb (Abschnitt 5.3.4)gewählt. Dies setzt die Erregungsmuster des geschätzten Sprach- und Störanteils ins Verhältnis.
- Die Filterregel WienerSNRbN2 (Abschnitt 5.4.3) verwendet für das Verhältnis die Lautheiten.

Die Betrachtungen werden analog zu den psychoakustischen Modellen wie in Abschnitt 6.3 für die Filterregeln durchgeführt.

6.4.2 Sprachverzerrung (cepstrale Distanz)

In Abbildung 6.3 soll der Einfluss der psychoakustischen Filterregeln als Teil der psychoakustischen Stufe (zweite Stufe) auf die Sprachverzerrung instrumentell mittels der

cepstralen Distanz aufgezeigt werden. Die Verläufe der konventionellen Störgeräuschreduktion mit Wiener Filter und der zweistufigen Störgeräuschreduktion mit den Filterregel HINDnEx sind nahezu identisch. Für letzteres liegt die cepstrale Distanz knapp unterhalb der des Wiener Filters. Lediglich für sehr niedrige Störabstäde von $< -8\,\mathrm{dB}$ verringert sich der Wert gegenüber dem Wiener Filter um 0.4 dB. Deutlich stärker unterscheidet sich die cepstrale Distanz von der des Wiener Filters beim Einsatz des lautheitsbasierten Wiener Filters WienerNSNRb2. Bei niedrigen Störabständen verringert dieser um ein halbes Dezibel, bei hohen Störabständen um ein Dezibel gegenüber der konventionellen Störgeräuschreduktion. Gerade bei hohen Störabständen wirkt sich das Verhältnis der Lautheiten des Sprach- und Störanteils positiv auf eine Verringerung der Sprachverzerrung aus. Hingegen liegt der Verlauf der cepstralen Distanz bei Nutzung des erregungsbasierten Wiener Filter WienerExSNRb2 für alle betrachteten Störabstände ca. 1 dB oberhalb des Verlaufs bei Nutzung des Wiener Filters.

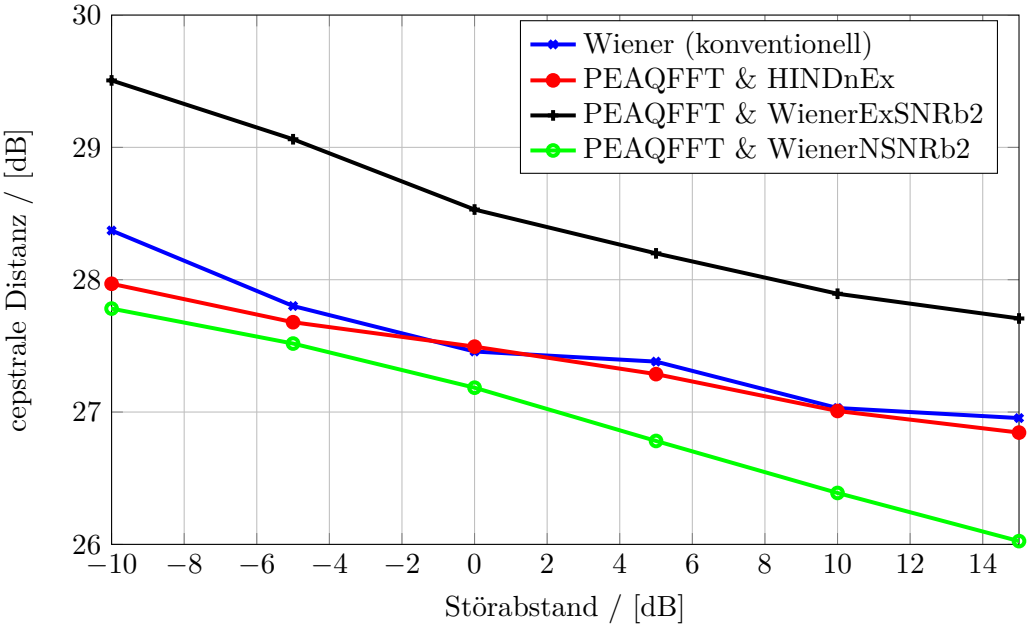


Abbildung 6.3: cepstrale Distanz in Abhängigkeit des Störabstands, konv. (Wiener) und zweistufige Störgeräuschreduktion mit den Filterregeln HINDnEx, WienerExSNRb2, WienerNSNRb2 unter Anwendung des psychoakustischen Modelle PEAQFFT

Beim Vergleich der Spektogramme des verbesserten Sprachsignals der Filter WienerExSNRb2 und WienerNSNRb2 (Abbildungen 6.5e, 6.5f) fällt jedoch auf, dass gerade beim hier bzgl. cepstraler Distanz schlecht abschneidenden erregungsbasierten Wiener Filter bei hohen Frequenzen noch mehr Sprachanteile vorhanden sind als beim lautheitsbasierten Wiener Filter. Für niedrige Frequenzen ($< 2\,\mathrm{kHz}$) ist die Amplitude des Sprachanteils sogar höher [3]. Allerdings ist der Rauschanteil, erkennbar durch den orangenen Bereich neben den hohen Sprachanteilen (rote Bereiche) deutlich größer als für den lautheitsbasierten Wiener Filter, wenn man die Spektogramme mit dem des Orignalsignals in Ab-

[3] rot entspricht einer hohen, blau einer niedrigen Amplitude des Spektrums des Signals

bildung 6.5a vergleicht. Daher muss bei der Berechnung der cepstralen Distanz für den erregungsbasierten Wiener Filter ein deutlich höherer Wert als für den lautheitsbasierten Wiener Filter herauskommen. Das heißt für die Zeitabschnitte hoher Amplitude im Sprachsignal ist das Filter WienerExSNRb2 bzgl. der Erhaltung der spektralen Anteile des Sprachsignals besser als das Filter WienerNSNRn2. In das instrumentelle Maß fließt jedoch auch das erhöhte Restrauschen ein. Im informellen Hörtest wird die Sprachverzerrung des erregungsbasierten Filters als geringer empfunden. Die cepstrale Distanz eignet sich daher zwar zur Bestimmung der Sprachverzerrung, jedoch weniger zur psychoakustisch empfundenen Sprachverzerrung.

Das lautheitsbasierte Wiener Filter WienerNSNRb2 weist mit der niedrigsten cepstralen Distanz die geringste Sprachverzerrung unter allen Filterregeln auf.

6.4.3 Störgeräuschreduktion - Differenz zwischen segmenteller Rausch- und Sprachdämpfung

Der Einfluss der Filterregeln der psychoakustischen Stufe der zweistufigen Störgeräuschreduktion auf die Differenz zwischen Störgeräusch- und Sprachdämpfung $SegDA$ist in Abbildung 6.4 zu sehen. Für sehr niedrige Störabstände -10 bis 8 dB führen die Filterregeln der zweistufigen Störgeräuschreduktion auf gleiche (WienerExSNRb2 und HINDnEx) oder höhere Werte (WienerNSNRb2). Für letzteren lässt sich dies mit dem Faktor $-1{,}5$ in der Filterregel 5.17 erklären, dessen Einfluss mit sinkendem Störabstand zunimmt. Die Filtergewichte werden dabei schneller als bei den anderen Filterregeln durch die Begrenzung auf das Intervall [0,1] auf Null gesetzt. Für höhere Störabstände (< 0 dB) steigt die Differenz zwischen Störgeräusch- und Sprachdämpfung nicht mehr an. Der oben genannte Faktor wirkt sich dann gegenüber der den anderen Filterregeln WienerExSNRb2 und HINDnEx (Gleichungen 5.3.4, 5.2) negativ aus, da er den Zähler des Gewichts und damit das Gewicht selbst stärker verkleinert.

Der erregungsbasierte Wiener Filter erreicht mit der zweistufigen Störgeräuschreduktion für hohe Störabstände eine wünschenswert hohe Differenz zwischen Störgeräusch- und Sprachdämpfung. Letztere korreliert zumindest für hohe Frequenzen stark mit der cepstralen Distanz, letzteres ein logarithmisches Maß ist. Von daher müsste man unter Einbeziehung der Abbildung 6.3 mit den vergleichsweise hohen Werten für cepstralen Distanz von einer niedrige Differenz der Dämpfungen ausgehen. Allerdings zeigt sich bei Betrachtung des Spektrogramms in Abbildung 6.5e, dass die Sprachdämpfung des Filters für hohe Frequenzen vergleichsweise gering ist, da die Sprachanteile bei Frequenzen um die 8 kHz verglichen mit dem lautheitsbasierten Wiener Filter relativ hoch sind. Das Spektogramm des mit letzterem gefilterten Sprachsignals weist kaum Anteile oberhalb 3 kHz auf.

Bei informellen Hörtest zeigt sich eine hörbar niedrigere Störgeräuschdämpfung des HINDnEx Filters. Daher fällt, auch bei im Vergleich ähnlicher hoher cepstraler Distanz (und damit korrelierenden Sprachdämpfung) wie bei der konventionellen Störgeräuschreduktion, die Differenz der Störgeräusch- und Sprachdämpfung niedriger aus. Der Verlauf liegt bis zu 4 dB unterhalb des Verlaufs des Wiener Filters. Da der HINDnEx Filter dem ursprünglichen HIND Filter nahezu gleicht, kann dies mit der oben genannten Erläuterung 6.3.3 zu letzterem erklärt werden.

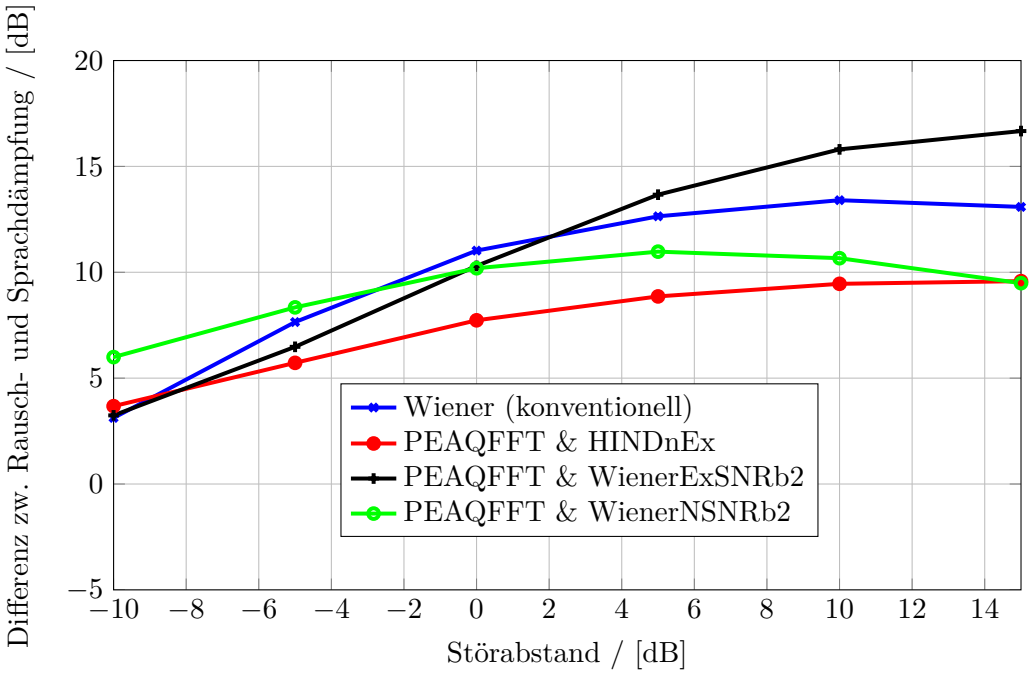


Abbildung 6.4: Differenz zwischen Rausch- und Sprachdämpfung in Abhängigkeit des Störabstands

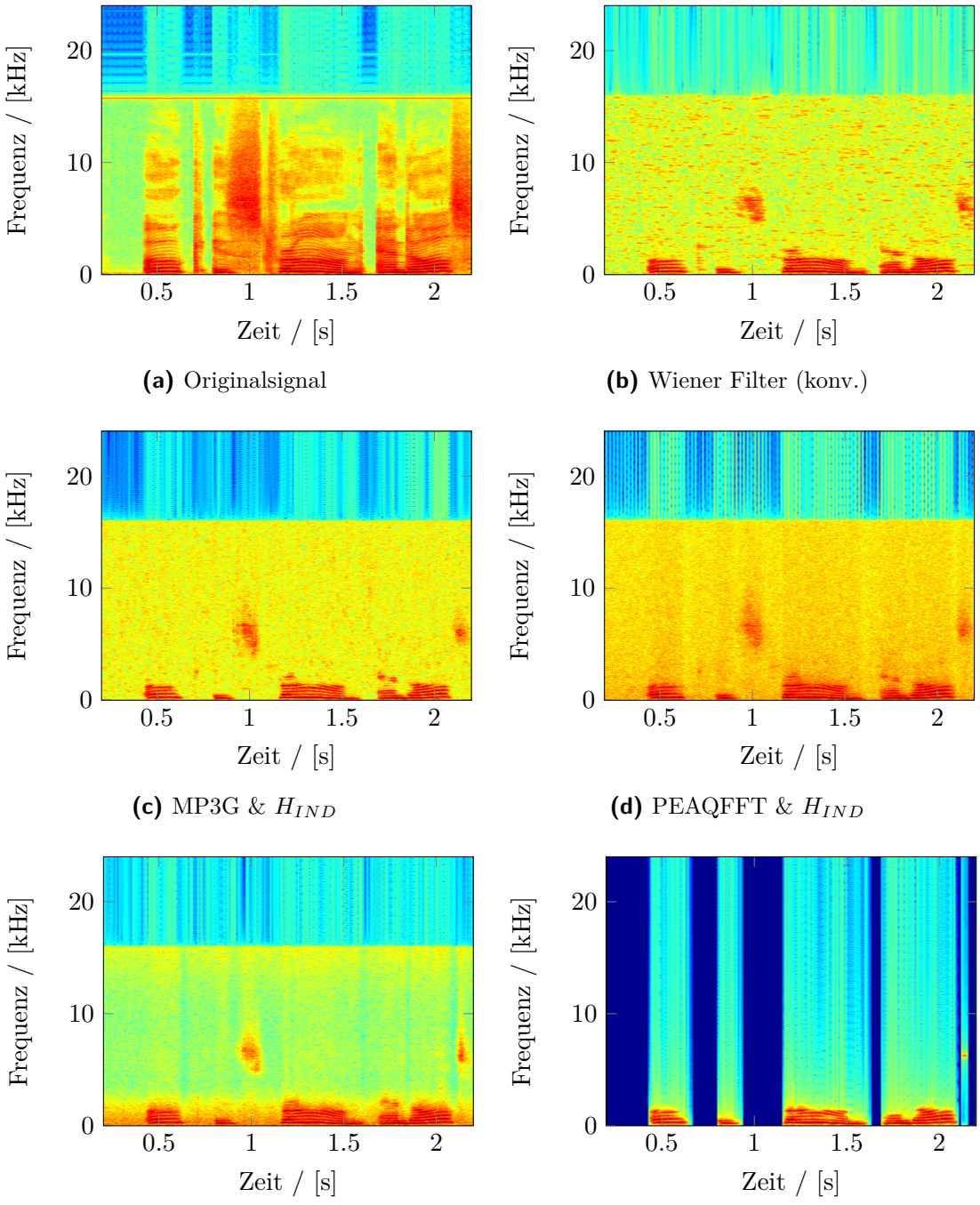
Zusammenfassend kann gesagt werden, dass der lautheitsbasierte Wiener Filter sich für niedrige Störabstände wegen des dabei hohen Verhältnisses von Störgeräusch- zu Sprachdämpfung eignet, während der erregungsbasierte Wiener Filter sich in dieser Hinsicht vorteilhaft bei hohen Störabständen auswirkt.

6.5 Musical Noise

Artefakte wie das „Musical Noise“ lassen sich mit keinem instrumentellen Maß messen. Zur Bewertung der behandelten Störgeräuschreduktionssysteme eignet sich daher nur der visuelle Eindruck des Spektrograms: „Musical Noise“ äußert sich durch auftretende „Flecken“ hoher Amplitude, welche die für dies charakteristischen vereinzelten Spitzen im Spektrum repräsentieren. Besonders in den hohen Frequenzbereichen mit typischerweise geringem Auftreten von Sprachanteilen sind die Spitzen gut erkennbar. Die Spektrogramme des Originalsignals und der verbesserten Signale der konventionellen und zweistufigen Störgeräuschreduktion sind in Abbildung 6.5 aufgetragen. Das verrauschte Signal wird durch Hinzufügen von weißem Rauschen generiert. Der Störabstand beträgt 5 dB. Für die zweistufige Störgeräuschreduktion dient das psychoakustische Modell MP3G in Kombination mit dem H_{IND} Filter nach [13] als Referenz für die Störgeräuschreduktionen, die das PEAQFFT Modell mit psychoakustischen Wiener Filterregeln, sowie dem H_{IND} Filter kombinieren. Die Filterregeln wurden in Abschnitt 6.4 miteinander verglichen.

Deutlich sind die vereinzelten Spitzen im Spektrogramm des verbesserten Signals der konventionellen Störgeräuschreduktion mittels Wiener Filters in der Unterabbildung 6.5b zu sehen. Das „Muscical Noise“ tritt an allen Stellen geringer Sprachamplitude (zeitlich

wie spektral betrachtet) auf.



(a) Originalsignal

(b) Wiener Filter (konv.)

(c) MP3G & H_{IND}

(d) PEAQFFT & H_{IND}

(e) PEAQFFT & WienerExSNRb2

(f) PEAQFFT & WienerNSNRb2

Abbildung 6.5: Spektogramme der originalen Sprachsignals a) und der verbesserten Sprachsignale b) bis f) des mit weißem Rauschen gestörten Signals, konventionelle Störgeräuschreduktion b) und zweistufige Störgeräuschreduktion mit verschiedenen Modellen und Filterregeln c) bis f)

Die von Gustafsson verwendete zweistufige Störgeräuschreduktion senkt die Amplituden dieser vereinzelten Spitzen im Spektrum deutlich ab (Abbildung 6.5c). Eine weitere Reduktion der Wahrnehmbarkeit erfolgt bei Nutzung des PEAQFFT Modelles in Verbindung mit dem H_{IND} Filter entsprechend Abbildung 6.5d. Allerdings ist bei dieser Variante auch der verbleibende Störanteil im verbesserten Signal höher. Die Wiener Filter basierten psychoakustischen Störgeräuschreduktionen in den letzten Abbildungen 6.5e und 6.5f weisen kaum für Musical Noise charakteristische Spitzen im Spektrum auf, obwohl die Reduktion des weißen Rauschens sehr stark ist – besonders für den lautheitsbasierten Wiener Filter. Alle Varianten der psychoakustisch basierten Störgeräuschreduktion erreichen eine signifikante Verringerung von „Musical Noise“.

6.6 Sprachverständlichkeitsmaß STOI

Die Erläuterungen der vorhergehenden Abschnitten dienen dazu die Störgeräuschreduktionssysteme bzgl. einzelner Charakteristiken wie Sprach- und Rauschdämpfung, Auftreten von *Musical Noise*, Sprachverzerrung etc. zu bewerten. Über diesen Aspekten steht die Sprachverständlichkeit, welche von den einzelnen Charakteristiken beeinflusst wird. Daher wird nun ganzheitlich die Sprachverständlichkeit mittels des von [35] dafür als am besten geeigneten Maß STOI evaluiert.

In der Abbildung 6.6 sind die Sprachverständlichkeitswerte der konventionellen Störgeräuschreduktion mittels Wiener Filter und der psychoakustisch basierten in Abhängigkeit des Störabstands aufgetragen.

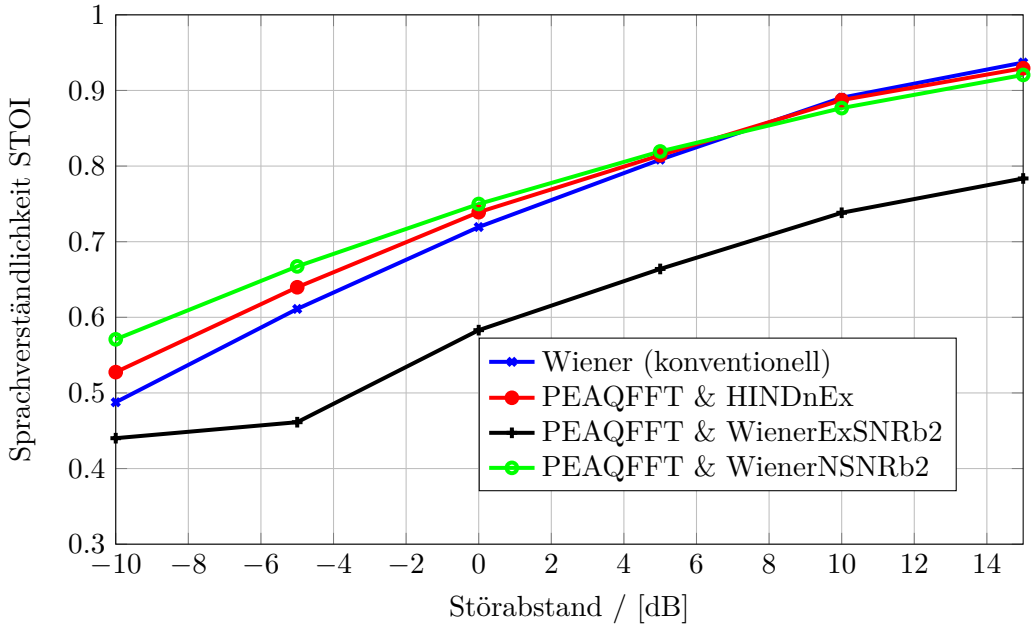


Abbildung 6.6: Sprachverständlichkeit der konventionellen und zweistufigen Störgeräuschreduktionen in Abhängigkeit des Störabstands

Die Sprachverständlichkeit ist allgemein für die psychoakustisch basierten Störgeräusch-

reduktionen für niedrige Störabstände bis 7 dB bis auf die mit erregungsbasiertem Wiener Filter (WienerExSNRb2) besser als bei der konventionellen Störgeräuschreduktion. Die Konvergenz der Sprachverständlichkeitswerte zum Wert des Wiener Filters für hohe Störabstände erklären sich durch die dabei höhere Genauigkeit der Schätzung des Störanteils. Somit wirken sich kompensatorische Glättungseffekte durch die Psychoakustik (Transformation in den Bereich kritischer Bänder, Spreizung etc.) nicht mehr so stark auf die Filtergewichte aus. Die in den Frequenzbereich zurücktransformierten Erregungsmuster der psychoakustischen Filterregeln nähern sich den Leistungsdichtespektren des Wiener Filters der konventionllen Störgeräuschreduktion an. Zudem ist der Rausch Maskierungsabstand NMR bei hohen Störabständen sehr groß, Verdeckungseffekte werden kaum noch ausgenutzt. Die schlechteren Werte des erregungsbasierten Wiener Filters (WienerNSNRb2) lassen sich durch die für diesen typischerweise hohe cepstrale Distanz wie in Abschnitt erklären. Die hohen Werte für die cepstrale Distanz werden unter anderem durch das relativ hohe Restrauschen im unteren Frequenzbereich hervorgerufen.

6.7 Zusammenfassung Ergebnisse

In diesem Kapitel konnte gezeigt werden, dass die entwickelten psychoakustischen Modelle und Filterregeln je nach Kombination eine Verringerung der Sprachverzerrung im Sinne der cepstralen Distanz erreichen. Darüber hinaus wird eine signifikante Reduktion von Artefakten wie Musical Noise erzielt. Je nach Filterregel liegt die Differenz der Rausch- zu Sprachdämpfung im Bereich der Werte für eine konventionelle Störgeräuschreduktion. Dieses Maß weist eine Korrelation zu der empfundenen Störgeräuschreduktion auf. Diese drei Kriterien werden allerdings abhängig von der Filterregel unterschiedlich erfüllt: Das HINDnEx Filter erreicht in der zweistufigen Störgeräuschreduktion bei der Sprachverzerrung ähnliche Werte wie das Wiener Filter in der konventionellen Störgeräuschreduktion. Gegenüber dem in dieser Hinsicht besseren lautheitsbasierten Filter WienerNSNRb2 zeigt sich beim Vergleich der Spektrogramme der verbesserten Sprachsignale, dass das HINDnEx Filter hochfrequente Sprachanteile deutlich weniger dämpft, allerdings auf Kosten einer geringeren Störgeräuschreduktion. Der erregungsbasierte Wiener Filter weist eine hohe Störgeräuschreduktion, aber auch eine zumindest instrumentell gemessen, hohe Sprachverzerrung auf, obwohl hochfrequente Sprachanteile ähnlich gut wie beim HINDnEx Filter bei der spektralen Gewichtung erhalten bleiben. Die Sprachverständlichkeit ist bei den zwei Filtervarianten HINDnEx und WienerNSNRb2 für niedrige Störabstände höher als bei der konventionellen Störgeräuschreduktion mittels Wiener Filter. Anzumerken ist, dass alle psychoakustischen Filter eine deutliche Reduktion von Musical Noise erzielen.

Kapitel 7

Zusammenfassung und Ausblick

In dieser Arbeit wird die konventionelle Störgeräuschreduktion um eine zweite Stufe erweitert, die aus einem psychoakustischen Modell und einer psychoakustischen Filterregel besteht und damit die Eigenschaften des menschlichen Gehörs inklusive Teile der neuronalen Ebene repräsentiert. Daher gliedert sich die Entwicklung der psychoakustisch basierten Störgeräuschreduktion in zwei Schritte: Optimierung und Erweiterung möglichst genauer psychoakustischer Modelle und anschließender Entwicklung von neuen Filterregeln.

Das Ziel und die Motivation für die Ausnutzung von Psychoakustik in der Störgeräuschreduktion liegen in der Verringerung der bei der konventionellen Störgeräuschreduktion auftretenden Sprachverzerrungen und Artefakte.

Bei konventionellen Methoden treten Fehler bei der Schätzung des Sprache und des Störanteils besonders stark bei niedrigen Störabständen auf. Diese treten vor allem im oberen Sprachband wegen der typischen spektralen Energieverteilung der Sprache auf, welche zu hohen Frequenzen abfallend ist. Durch Fehlerfortpflanzung vergrößert sich der Fehler im resultierenden Filtergewicht. Dies führt zu dem sogenannten „Musical Noise“, welches sich akustisch durch ein zeitlich schnell veränderndes Hintergrundgeräusch äußert. Ein weiteres Problem der Störgeräuschreduktion liegt in der teilweise starken Dämpfung des Sprachanteils, wenn eine gewisse Störgeräuschreduktion erzielt werden soll.

Mittels der zweistufigen Störgeräuschreduktion wird zunächst unter Ausnutzung spektraler und temporaler Verdeckungseffekte der im verrauschten Signal vorhandene Störanteil nur soweit reduziert, bis dieser nicht mehr wahrgenommen werden kann. Daraus ergibt sich eine objektiv verringerte Störgeräuschreduktion, die theoretisch subjektiv nicht wahrgenommen wird.

Die für die Berechnung der Verdeckung nötigen psychoakustischen Modelle werden im ersten Schritt auf die Leistungsfähigkeit, die Verarbeitung des Signals durch das menschliche Gehör nachzubilden, untersucht. Darüber hinaus werden die Modelle um fehlende Eigenschaften erweitert und für den Einsatz in einer Störgeräuschreduktion angepasst. Im zweiten Schritt wird die Filterregel, welche Verdeckungseffekte nutzt, durch eine allgemein psychoakustisch basierte Filterregel abstrahiert. Dies führt auf zwei neue Filterkatergorien: erregungsbasierte und lautheitsbasierte Wiener Filter. Die Optimierung dieser Regeln findet unter Ausführung zahlreicher informeller Hörtests statt. Dabei werden die Filterregeln mit dem erweiterten FFT-basierten psychoakustischen Modell [37] kombiniert, wel-

ches sich im ersten Schritt für die Berechnung der verdeckungsbasierten Filterregel als das beste erwiesen hat. Das verdeckungsbasierte Filter, wie auch die besten Varianten der neuen Filterkategorien haben in zahlreichen informellen Hörversuchen und Betrachtungen von Spektrogrammen gezeigt, dass Musical Noise sowohl für transiente wie auch stationäre Störsignale unterdrückt wird. Das verdeckungsbasierte Filter verursacht subjektiv gesehen die geringste Sprachverzerrung über dem gesamten Frequenzbereich. Allerdings fällt die Störgeräuschreduktion geringer aus als bei Anwendung der konventionellen Störgeräuschreduktion. Das erregungsbasierte Wiener Filter weist eine höhere Störgeräuschdämpfung hoher Frequenzen als das verdeckungsbasierte Filter auf. Jedoch ist der Anteil von Restrauschen bei niedrigen Frequenzen relativ hoch. Eine höhere Störgeräuschdämpfung als die konventionelle Störgeräuschreduktion wird mit dem lautheitsbasierten Wiener-Filter erreicht. Bei einem Störabstand von 5 dB wird der Störanteil vollständig unterdrückt. Beim Betrachten der Sprachverzerrung weist das verdeckungsbasierte Filter und das lautheitsbasierte Wiener- Filter gemessen an der cepstralen Distanz für Störabstände von -10 bis 15 dB durchgehend niedrigere Werte gegenüber der konventionellen Störgeräuschreduktion auf. Für hohe Störabstände ist die cepstrale Distanz für den lautheitsbasierten Filter um bis zu 1 dB abgesenkt. Dieser reduziert allerdings hochfrequente Signalanteile stärker als die konventionelle Störgeräuschreduktion.

Die Ausnutzung der Psychoakustik ist durch die spektrale Energieverteilung der Sprache begrenzt. Jedoch erreicht die psychoakustisch basierte Störgeräuschreduktion eine höhere Störgeräuschdämpfung bei gegenüber der konventionellen Störgeräuschreduktion gleichbleibender Sprachdämpfung.

7.1 Ausblick

In Kapitel 4 wird das Filterbank basierte Modell untersucht, welches aufgrund der einfachen Rücktransformation nicht die erwartete Verbesserung gegenüber dem PEAQFFT Modell bietet. Eine verbesserte Rücktransformation könnte die Genauigkeit der psychoakustischen Repräsentation in der zweistufigen Störgeräuschreduktion verbessern.

Da das lautheitsbasierte Filter die höchste Störgeräuschreduktion aufweist, allerdings die hohen Frequenzen stark unterdrückt werden, könnte die spektrale Korrelation des unteren und oberen Sprachbands genutzt werden, um mittels einer Bandbreitenerweiterung den Höreindruck zu verbessern.

Die Störgeräuschreduktion könnte dem Gehör nachempfunden werden, in dem die zweistufige Störgeräuschreduktion komplett im Bereich der kritischen Bänder durchgeführt wird. Damit ließen sich ggf. die Auswirkungen der Fehler durch Transformationen minimieren.

Da die psychoakustischen Filterregeln teilweise für unterschiedliche Bereiche des Störbabstands bessere oder schlechtere Leistungen aufweisen, liegt es nahe zwei, Modelle parallel, und je nach Störabstand das Ergebnis der ein oder anderen Filterregel zu verwenden.

Alternative Lösungen liegen im Bereich der Verbesserung der Schätzung. Mit der statistischen Verteilung der Schätzfehler in Abhängigkeit des Störabstands könnte ein Gewichtungsfaktor bestimmt werden, der das spektrale Gewicht in Abhängigkeit der Genauigkeit nachträglich gewichtet. Die Verwerfung des Störabstands bei Unterschreitung eines Schwellwerts wie in Kapitel 5 hat gezeigt, dass dies je nach Schwellwerthöhe zu besseren

Ergebnissen führt, als mit fehlerhaften Störabständen Filtergewichte zu bestimmen.

Literaturverzeichnis

[1] ITU recommendations bs 1387: method for objective measurements of subjectivley perceived audio quality.

[2] Brian C. J. , Brian R. Glasberg, and Thomas Baer. A model for the prediction of thresholds, loudness, and partial loudness. *J. Audio Eng. Soc*, 45(4):224–240, 1997.

[3] H. Abdolvahab-Emminger and C. Benz. *Physikum exakt: das gesamte Prüfungswissen für die 1. ÄP ; 199 Tabellen ; [ideal für die neue AO].* Thieme, 2005.

[4] A.Oxenham. Forward masking: adaption or integration? *Acoustical Society of America, vol. 109, no2, pp.732-741*, 2001.

[5] V. Atti A.Spanias, T. Painter. *Audio signal processing and coding.* Wiley, 2007.

[6] Richard H. Ehmer. Masking by tones vs noise bands. *The Journal of the Acoustical Society of America*, 31(9), 1959.

[7] Thomas Esch. *Model-Based Speech Enhancement Exploiting Temporal and Spectral Dependencies.* Dissertation, IND, RWTH Aachen, April 2012.

[8] T. Gerkmann. Improved MMSE – based noise PSD tracking using temporal Cepstrum smoothing. *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, NY, USA*, page 1, October 2011.

[9] Timo Gerkmann and Richard C. Hendriks. Noise power estimation based on the probability of speech presence. In *WASPAA*, pages 145–148. IEEE, 2011.

[10] BR Glasberg and BC Moore. Derivation of auditory filter shapes from notched-noise data. *Hearing Research*, 47(1-2):103–138, 1990.

[11] Ben Gold, Nelson Morgan, and Dan Ellis. *Speech and Audio Signal Processing - Processing and Perception of Speech and Music, Second Edition.* Wiley, 2011.

[12] Teddy Surya Gunawan. *Audio Compression and Speech Enhancement using Temporal Masking Models.* PhD thesis, School of Electrical Engineering and Telecommunications - The University of New South Wales, Jan 2007.

[13] S. Gustafsson. Enhancement of Audio Signals by Combined Acoustic Echo Cancellation and Noise Reduction. *Aachener Beiträge zu Digitalen Nachrichtensystemen*, March 1999.

[14] E. Zwicker H. Fastl. *Psychoacoustics, Facts and Models.* Springer, Berlin, Germany, 2007.

[15] J.L. Hall. *Auditory psychophysics for coding applications, in The Digital Signal Processing Handbook, V. Madisetti and D. Williams, Eds. Boca Baton.* Fl: CRC Press, 1998.

[16] Florian Heese. Modellbasierte Störgeräuschreduktion für breitbandige Sprache. Diplomarbeit, IND, RWTH Aachen, April 2009.

[17] RhonaP. Hellman. Asymmetry of masking between noise and tone. *Perception And Psychophysics*, 11(3):241–246.

[18] International Organization for Standardization (ISO). ISO Standard 226:2003-3. Geneva, 2003.

[19] ISO/IEC. Iso/iec 11172-3:1993 - information technology – coding of moving pictures and associated audio for digital storage media at up to about 1,5 mbit/s – part 3: Audio. Padrão, 1993.

[20] M. Schoenwiesner. J. Timoney, T. Lysaght. Implementing loudness models in matlab. In *7th Int. Conference on Digital Audio Effects (DAFx'04)*, Naples, Italy, Oct 2004.

[21] A.Spanias Jayaraman J. Thiagarajan. Morgan & Claypool, Arizona State University, 2012.

[22] M. Karjalainen. Auditory models for speech processing. volume 2, pages 11–20.

[23] K. Kroschel K.D. Kammeyer. Springer Vieweg, Wiesbaden, Germany, 2012.

[24] R.A. Lutfi. Additivity of simultaneous masking. pages 262–267, 1983.

[25] M. Vorländer. Technische Akustik I und II, Unterlagen zur Vorlesung, RWTH Aachen, Lehrstuhl für Technische Akustik, 2005.

[26] B.J.C. Moore. Academic Press, 2003.

[27] Brian C. J. Moore. Characterization of simultaneous, forward and backward masking. In *Audio Engineering Society Conference: 12th International Conference: The Perception of Reproduced Sound*, Jun 1993.

[28] J.L. Hall M.R.Schroeder, B.S. Atal. Optimizing digital speech coders by exploiting masking properties of the human ear. *Acoustic Society of America, 66(6)*, 1979.

[29] Hossein Najaf-Zadeh, Hassan Lahdili, Louis Thibault, and Michel C. Lavoie. Use of auditory temporal masking in the mpeg psychoacoustic model 2. In *Audio Engineering Society Convention 114*, Mar 2003.

[30] E. Shriberg O. Cetin. Speaker Overlaps and ASR errors in meetings: Effects before, during, and after the overlap. *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2006.

[31] R. Martin P. Vary. *Digital Speech Transmission - Enhancement Coding and Error Concealment*. Wiley, Chichester, England, 2006.

[32] T. Painter and A. Spanias. Perceptual coding of digital audio. April 2000.

[33] R. D. Patterson, K. Robinson, J. Holdsworth, D. Mckeown, C. Zhang, and M. Allerhand. Complex sounds and auditory images. In *in Proc. 9th Int. Symp. Hearing Audit., Physiol. Perception*, pages 429–446, 1992.

[34] HEAD acoustics GmbH Prof. Dr.-Ing. K. Genuit. Vorlesung Psychoakustik. Institut für technische Akustik - RWTH AACHEN, October 2012.

[35] Taal CH, Hendriks RC, Heusdens R, Jensen J., ITG Fachtagung Sprachkommunikation. Intelligibility prediction of single-channel noise-reduced speech. Delft, 2010.

[36] Thilo Thiede. Eine rechenzeiteffektive gehörrichtige filterbank mit signalabhängiger filtercharakteristik. *Impulse und Antworten - Festschrift für Manfred Krause*, pages 263–272, 1999.

[37] Thilo Thiede, William C. Treurniet, Roland Bitto, Christian Schmidmer, Thomas Sporer, John G. Beerends, and Catherine Colomes. PEAQ - The ITU Standard for Objective Measurement of Perceived Audio Quality. *J. Audio Eng. Soc*, 48(1/2):3–29, 2000.

[38] Thilo Thiede, William C. Treurniet, Roland Bitto, Thomas Sporer, Karlheinz Brandenburg, Christian Schmidmer, Michael Keyhl, John G. Beerends, Catherine Colomes, Gerhard Stoll, and Bernhard Feiten. PEAQ - der künftige ITU-Standard zur objektiven Messung der wahrgenommenen Audioqualität. In *20. Tonmeistertagung*, Karlsruhe, 1998.

[39] Hartmut Traunmüller. Analytical expressions for the tonotopic sensory scale. *The Journal of the Acoustical Society of America*, 88(1), 1990.

[40] E. Zwicker and A. Jaroszewski. Inverse frequency dependence of simultaneous tone-on-tone masking patterns at low levels. *Journal of the Acoustical Society of America*, 71(6):1508–1512, 1982.

[41] E. Zwicker and E. Terhardt. Analytical expressions for criticalband rate and critical bandwidth as a function of frequency. *The Journal of the Acoustical Society of America*, 68(5), 1980.

[42] Eberhard Zwicker, Hugo Fastl, Ulrich Widmann, Kenji Kurakata, Sonoko Kuwano, and Swiichiro Namba. Program for calculating loudness according to din 45631 (iso 532b). *Journal of the Acoustical Society of Japan (E)*, 12(1):39–42, 1991.

[43] Eberhard Zwicker and Richard Feldtkeller. *Das Ohr als Nachrichtenempfänger*. Hirzel, 2., neubearb. Aufl. edition, 1967.