Friedrich-Alexander-Universität Erlangen-Nürnberg

Lehrstuhl für Multimediakommunikation und Signalverarbeitung

Prof. Dr.-Ing. Walter Kellermann

Bachelorarbeit

Analyse von psychoakustischen Modellen für einkanalige Störreduktionsverfahren

von Lena Badstieber

September 2013

Betreuer: Dipl.-Ing. Klaus Reindl
medizinischer Betreuer: Dr. Ulrich Hoppe

Erklärung

Ich versichere, dass ich die vorliegende Arbeit ohne fremde Hilfe und ohne Benutzung anderer als der angegebenen Quellen angefertigt habe, und dass die Arbeit in gleicher oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegen hat und von dieser als Teil einer Prüfungsleistung angenommen wurde. Alle Ausführungen, die wörtlich oder sinngemäß übernommen wurden, sind als solche gekennzeichnet.

_____ _____

Ort, Datum Unterschrift

Inhaltsverzeichnis

Kurzfassung

Hörgeräteträgern fällt es oft schwer sich in Situationen, in denen viele Geräusche aktiv sind, gut zu verständigen. Daher sind Algorithmen zur Störreduktion notwendig, die diese Situation verbessern können, indem sie die Störquellen unterdrücken. Allerdings haben diese Algorithmen den Nachteil, dass besonders in niedrigen SNR Bereichen (< 10 dB) Artefakte und *musical tones* auftreten. Die einkanaligen Filter die zur Störreduktion verwendet werden, sollen in dieser Arbeit durch Parameter optimiert werden. Die Parameter zur Störreduktion werden anhand der Maskierung des menschlichen Gehörs optimiert. Die Maskierung ist eine Eigenschaft, bei der ein schwacher Ton durch einen dominanten Ton verdeckt wird. Anhand dieser Verdeckungseffekte sollen die Artefakte verdeckt und somit nicht hörbar gemacht werden. Der Schwerpunkt liegt zum einen auf der Berechnung des Masking Thresholds durch den diese Eigenschaft nachgebildet werden kann und zum anderen auf der Berechnung der optimalen Parameterwerte, die aus diesem Threshold abgeleitet werden.

Kapitel 1

Einleitung

Viele Menschen kennen die Problematik des Verstehens, das Gefühl, dass die Mitmenschen undeutlich sprechen oder, dass in Situationen mit vielen Personen, die gleichzeitig reden, die Stimme des Gesprächspartners unter all den anderen verschwindet und so zu Verständnisproblemen führt. Dies sind oft erste Anzeichen einer Hörschwäche und ein Hörgerät wird womöglich nötig sein. Hier gibt es eine große Auswahl an Modellen, sodass jeder etwas Passendes finden kann. Allerdings ist es auch mit einem Hörgerät nicht einfach in Umgebungen, in denen viele Geräuschquellen gleichzeitig aktiv sind, den Gesprächspartner gut zu verstehen. Verfahren, die verwendet werden um Gespräche in solchen Situationen zu ermöglichen, basieren auf der Hervorhebung der gewünschten Quelle, indem die Störquellen unterdrückt werden. Dazu gibt es verschiedene Möglichkeiten, bei denen allerdings durch die Rauschunterdrückung störende Geräusche und Artefakte erzeugt werden. Diese führen zu einer erneuten Störung des Signals und in sehr verrauschten Situationen ist es sogar möglich, dass das erzeugte Signal schlechter ist als das Ursprüngliche. An genau diesem Punkt setzt die Idee dieser Arbeit an, denn es sollen die Eigenschaften des menschlichen Gehörs genutzt werden um die Störquellen auf natürliche Art und Weise für den Menschen unhörbar zu machen und somit zusätzliche Artefakte und Störungen zu verringern. In den folgenden Kapiteln werden zunächst die Eigenschaften des menschlichen Gehörs beschrieben und die einkanaligen Verfahren, die zur Störreduktion verwendet werden, erklärt. Danach wird erläutert wie

diese in der Arbeit genutzt und nachimplementiert werden. Abschließend erfolgt eine Auswertung der Ergebnisse.

Kapitel 2

Grundlagen des menschlichen Gehörs

In diesem Kapitel geht es um die Anatomie und die Eigenschaften des menschlichen Gehörs. Diese sind eine wichtige Voraussetzung zur Optimierung von Störreduktionsverfahren bei Hörgeräteanwendungen. Deshalb werden zu Beginn die wichtigsten Eigenschaften des menschlichen Ohrs näher erklärt.

2.1 Anatomie des menschlichen Gehörs

Das Kapitel basiert auf den Informationen aus [1, Kapitel 2.4]. Das menschliche Ohr ist in Abb. 2.1 dargestellt und gliedert sich grob in drei Teile. Es handelt sich dabei um das Außenohr, das Mittelohr und das Innenohr, die in Abb. 2.1 zu sehen sind.
Das Außenohr: Dies ist der größte und einzige von außen sichtbare Teil des Ohrs. Es besteht aus der Ohrmuschel, die das Ohr schützt und zusammen mit dem Kopf und den Schultern den Einfallsschall codiert. Ein weiterer Bestandteil des Außenohrs ist der äußere Ohrkanal, der eine gleichmäßige Röhre mit einem Durchmesser von 0.7 cm und einer Länge von 3 cm bildet. In Abb. 2.1 ist dieser direkt im Anschluss an die Ohrmuschel zu finden. Die Aufgabe des äußeren Ohrkanals ist es den Schall zum Trommelfell zu leiten. Da dessen Resonanzfrequenz zwischen 3kHz und 4kHz liegt,

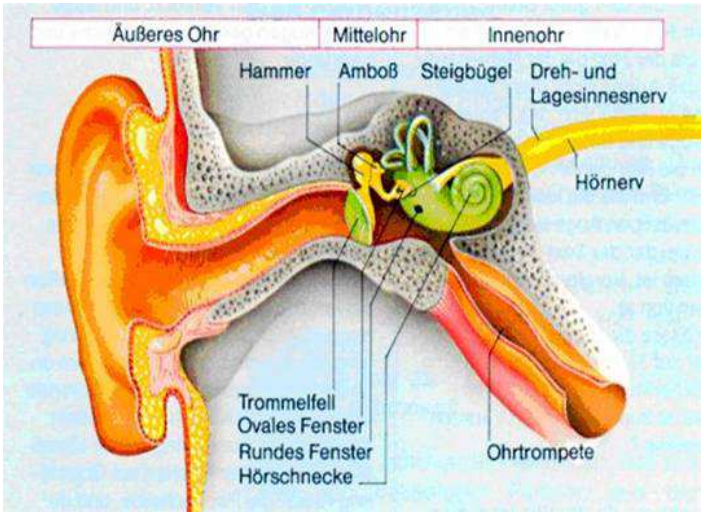


Abbildung 2.1: Übersicht des menschlichen Gehörs [2]

ist unser Ohr in diesem Frequenzbereich besonders empfindlich. Den letzten Teil des Außenohrs bildet das Trommelfell. Es stellt eine steife, zapfenförmige Membran dar, die aufgrund von Kräften der schwingenden Luftpartikel vibriert. In der Abb. 2.1 ist es der kleine hellgrüne Bereich am Ende des äußeren Ohrkanals.

Das Mittelohr: Das Mittelohr ist ein luftgefüllter Hohlraum, der auf der einen Seite durch das Trommelfell mit dem Außenohr und auf der anderen Seite durch das ovale und das runde Fenster, die in Abb. 2.1 zu sehen sind, mit dem Innenohr verbunden ist. Es besteht aus den 3 kleinen Knochen Hammer, Amboss und Steigbügel. Diese werden gemeinsam als Gehörknöchelchen bezeichnet. Sie stellen eine akustische Verbindung zwischen dem Trommelfell und dem ovalen Fenster her. Aufgrund des Flächenverhältnisses von Trommelfell zu ovalem Fenster erfolgt eine Impedanzwandlung mit dem Faktor 15 vom luftübertragenen Schall zu der Flüssigkeit im Innenohr. Das bedeutet: Schwingungen der Luftpartikel, die kleine Kräfte und eine große Auslenkung besitzen, werden zu Schwingungen mit großen Kräften und kleiner Auslenkung transformiert. Außerdem besitzt das Mittelohr noch eine zusätzliche Verbindung, die Eustachische Röhre, auch Ohrtrompete genannt, die das Mittelohr mit dem Nasenrachenraum verbindet. Sie ist in der Abb. 2.1 am rechten unteren Ende in orange zu sehen. Sie verbindet das Mittelohr mit dem Nasenrachenraum und ist während des Schluckens geöffnet

um so den Druck im Mittelohr auszugleichen. Dies ist wichtig um den Ruhepunkt des Trommelfells mit dem Angriffspunkt der Gehörknöchelchen abzustimmen.

Das Innenohr: Im Innenohr befinden sich sowohl das Gleichgewichtsorgan als auch das Organ zur Orientierung, das aus halbkreisförmigen Kanälen besteht. Weiterhin gibt es noch die Cochlea, auch Gehörschnecke genannt, an deren Enden sich das runde und das ovale Fenster befinden. Sie ist wie ein Schneckenhaus geformt mit 2.5 Bögen und ist aufgerollt 32 mm lang. In der Abb. 2.1 ist die Cochlea gut zu erkennen, da die Schneckenhausform bildlich angedeutet ist. Die Gehörknöchelchen, in Abb. 2.1 direkt vor der Cochlea gelb markiert, übertragen die Bewegungen der Luftpartikel durch das ovale Fenster auf die Flüssigkeit der Basilarmembran, welche in Abb. 2.2 als eine dünne rote Linie zwischen dem Scala media und dem Scala tympani zu finden ist.

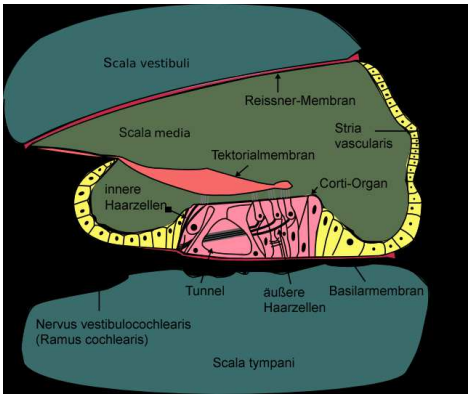


Abbildung 2.2: Aufbau der Cochlea [3]

Diese bildet einen Teil der Cochlea und trennt diese in zwei Räume. Die Aufteilung ist in Abb. 2.2 veranschaulicht, welche einen Schnitt durch die Cochlea darstellt. Hierbei bildet den oberen Teil, das *Scala vestibuli*, und den unteren, das *Scala tympani*. Diese beiden Räume sind an der Spitze, über das *Helicotrema*, miteinander verbunden. Auf der Basilarmembran befindet sich das Corti-Organ, das als Schnittstelle zwischen den akustisch-mechanischen Schwingungen und den Nervensignalen dient. In der Abb. 2.2 in rosa dargestellt. Es nimmt mit seinen 3600 inneren und 2600 äußeren Haarzellen die Schwingungen auf und gibt die Informationen an den Hörnerv und über neuronale Synapsen an das Gehirn weiter.

Durch das Helicotrema und das runde Fenster erfolgt ein Druckausgleich der Wanderwelle. In Abb. 2.3 ist eine solche Wanderwelle des menschlichen Gehörs abgebildet. Sie entsteht auf der Basilarmembran, welche am ovalen Fenster 0.05 mm, am Helicotrema hingegen 0.5 mm breit ist. Die Basilarmembran transformiert mittels eines Wanderwellenmechanismus eine bestimmte Schallfrequenz auf einen bestimmten Ort auf der Basilarmembran.

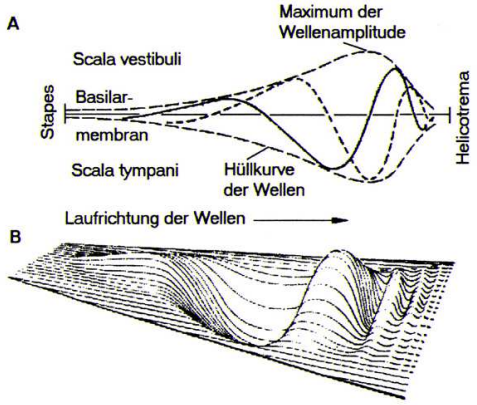


Abbildung 2.3: Fortbewegung der Wanderwelle auf der Basilarmembran [4]

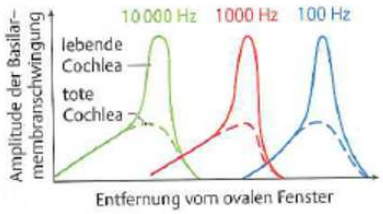


Abbildung 2.4: Transformation der Frequenzen auf die Basilarmembran [5]

In der Abb. 2.4 sind die Amplituden der Basilarmembranschwingungen in Bezug auf die Entfernung vom ovalen Fenster aufgetragen. Aus dieser lässt sich erkennen, dass die Basilarmembran und die Haarzellen des Corti-Organs durch hohe Frequenzen nahe des ovalen Fensters angeregt werden. Niedrige Frequenzen hingegen haben ihre Resonanz beim Helicotrema, welches sich in Abb. 2.4 an der Stelle mit der weitesten Entfernung

vom ovalen Fenster befindet. So entspricht jeder Ort auf der Basilarmembran einer bestimmten Frequenz, bei welcher dieser Teil mit maximaler Auslenkung schwingt. Diese Transformation korrespondiert mit einer Spektralanalyse, einer (nicht-gleichmäßigen) Filterbank. Die Hüllkurve der Auslenkung ist in Richtung des ovalen Fensters flach, beim Helicotrema jedoch sehr steil. Dieses Verhalten wird durch die Eigenschaften der Maskierung, die in Kapitel 2.2.3 noch näher erläutert wird, bestimmt.

2.2 Psychoakustische Eigenschaften

2.2.1 Hörvermögen und Lautstärke

Als Quelle diente für diesen Abschnitt [1, Kapitel 2.5.1]. Die menschliche Sprache entsteht durch kleine zeitliche Änderungen des Schalldrucks $p(t)$, der in Pascal [Pa] angegeben wird. Der Mensch kann einen sehr großen Tonumfang wahrnehmen. Dieser umfasst sieben Dekaden, wobei die unterste, bei der sogenannten Hörschwelle, bei 10^{-5} Pa beginnt und bei der Schmerzgrenze von 10^2 Pa endet. Der normierte Druck und die Intensität werden auf einer logarithmischen Skala gemessen. Der Schalldruckpegel ist durch folgende Gleichung gegeben:

$$L = 20 \cdot \log_{10}(\frac{p}{p_0}) = 10 \cdot \log_{10}(\frac{I}{I_0}) \tag{2.1}$$

Der Schalldruckpegel wird in Dezibel (dB) angegeben. Der Referenzschalldruck beträgt $p_0 = 20\mu$ Pa und die dazugehörige Schallintensität liegt bei $I_0 = 10^{-12}\frac{W}{m^2}$.
Die Hörfläche, die in Abb. 2.5 dargestellt ist, veranschaulicht die hörbaren Frequenzbereiche des menschlichen Gehörs. Auf dieser sind verschiedene Bereiche mit ihren dazugehörigen Frequenzen und Lautstärkepegeln zu sehen. Der Bereich den der Mensch wahrnehmen kann beginnt oberhalb der Hörschwelle und endet bei der Linie der Schmerzempfindung im oberen Bereich der Abb. 2.5. Der Threshold in Ruhe, der unterhalb der Hörschwelle in Abb. 2.5 liegt, ist frequenzabhängig. Die höchste Empfindlichkeit des Gehörs liegt zwischen 3 und 4 kHz, was durch die Resonanzfrequenz des Außenohrs

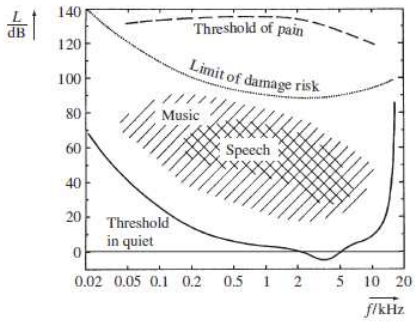


Abbildung 2.5: Hörfläche des Menschen [1, Kapitel 2.5.1]

bedingt wird, wie in Kapitel 2.1 beschrieben.

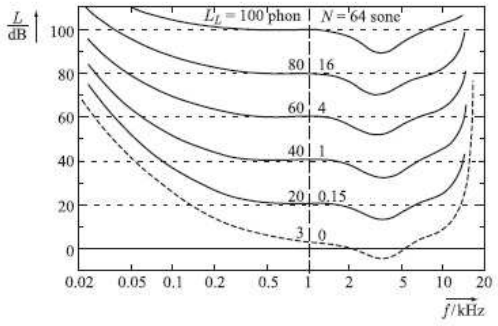


Abbildung 2.6: Grenzen des Lautstärkepegels [1, Kapitel 2.5.1]

Die Grenzen der konstanten subjektiven Lautstärke können durch einen auditiven Vergleich verschiedener sinusförmiger Testtöne bei verschiedenen Frequenzen und Amplituden und einer Referenz bei 1 kHz gefunden werden. Diese Grenzen sind in Abb. 2.6 bildlich dargestellt. Am Punkt der Referenzfrequenz gilt, dass der Lautstärkepegel und der Schalldruck identisch sind. Das bedeutet, dass der Lautstärkepegel L_L eines Testtons dem Lautstärkepegel der Referenzfrequenz mit Schalldruck $L = L_L$ gleicht. Der Lautstärkepegel wird in einer Pseudo-Einheit, Phon, angegeben. Die wahrgenommene Laustärke wird durch Hörtests ausgewertet und in der Einheit Sone angegeben. Eine Beziehung zwischen der Lautstärke N in Sone und dem Lautstärkepegel L_L in Phon gibt die folgende Gleichung:

$$N \approx 2^{(L_L - 40)/10} \tag{2.2}$$

Aus dieser Gleichung geht hervor, dass eine Laustärke $N = 1$ Sone einem Laustärkepegel $L_L = 40$ Phon entspricht.

2.2.2 Spektrale Auflösung

Die Informationen für diesen Abschnitt stammen aus [1, Kapitel 2.5.2]. Die spektrale Auflösung steht in enger Relation zur Frequenz-Ort Transformation der Basilarmembran. Daher hat der Mensch eine hohe spektrale Auflösung, da die Cochlea Sinneszellen für viele Frequenzen hat. Dennoch kann der Mensch nur begrenzt Unterschiede zwischen verschiedenen Tönen wahrnehmen. Ein Trägersignal mit Frequenz f_0 und ein frequenzmoduliertes Signal, welches dann seine spektralen Hauptkomponenten bei $f_0 \pm \Delta f$ besitzt, benötigt eine gewisse Abweichung, damit der Mensch den Unterschied zwischen den Signalen überhaupt wahrnimmt. Die gesamte Abweichung liegt in einem Bereich von $2 \cdot \Delta f$. So gilt, dass der Mensch bei Frequenzen kleiner als 500 Hz eine Abweichung von $2 \cdot \Delta f = 3.6$ Hz wahrnehmen kann, bei Frequenzen über 500 Hz erhöht sich der Wert proportional und es gilt:

$$2 \cdot \Delta f \approx 0.007 \cdot f_0 \tag{2.3}$$

Ein weiterer wichtiger Punkt ist die effektive spektrale Auflösung. Diese ist für unsere Auffassung der Lautstärke verantwortlich. Wenn ein Testsignal genutzt wird, das aus sinusförmigen Tönen oder einem Bandpass Rauschen mit anpassungsfähiger Bandbreite besteht, kann gezeigt werden, dass das menschliche Ohr die Reizung (Erregung) über bestimmte Frequenzintervalle integriert. Dadurch können im Bereich von 0 Hz bis 16 kHz 24 Intervalle definiert werden. Diese Intervalle werden als *critical bands* bezeichnet. Es besteht eine starke Korrelation zwischen den *critical bands* und dem Reizmuster (Stimulusmuster) der Basilarmembran. Wenn nun die Länge der Basilarmembran ebenfalls in 24 gleichgroße Intervalle geteilt wird entstehen 24 Segmente, die jeweils einem *critical band* entsprechen. Man nennt diese gleichmäßige Einteilung *bark scale*. Diese besonderen Merkmale der *critical bands* können genutzt werden um die Komplexität der Sprache zu verringern.

2.2.3 Maskierung

Das Wissen basiert auf [1, Kapitel 2.5.3]. Unter Maskierung wird das Prinzip verstanden, bei dem ein dominanter Ton einen schwächeren Ton verdeckt. Sie ist jedem einzelnen von uns bekannt, denn sie tritt nahezu täglich, bei fast allen Menschen auf. Eine solche Situation ist beispielsweise gegeben, wenn in einer lauten Umgebung Musik läuft und dadurch das Klingeln des Handys überhört wird. Dass das Klingeln nicht wahrgenommen wird, liegt an der Eigenschaft des menschlichen Gehörs, den schwächeren Ton zu verdecken. Man bezeichnet daher den dominanten Ton als Masker und den schwächeren als Testton. Im Folgenden wird dies noch genauer an der Abb. 2.7 erläutert: Der Masker liegt in diesem Fall bei f_M= 1kHz mit einer Bandbreite von Δf_M = 160 Hz und einem festen Schalldruckpegel L_M. Für jeden beliebigen Ton mit einer Frequenz f_T gilt, dass er maskiert wird, solange sein Spektrum L_T unterhalb des entsprechenden *masking thresholds* liegt. Jede Kurve in diesem Bild entspricht einem solchen *masking threshold*. Dass die Kurven nicht symmetrisch, sondern auf der Seite der niederen Frequenzen steiler sind, liegt an der fortschreitenden Welle auf der Basilarmembran. Die Senken bei der 80 dB und der 100 dB Kurve treten aufgrund von nicht-linearen Effekten des menschlichen Gehörs auf.

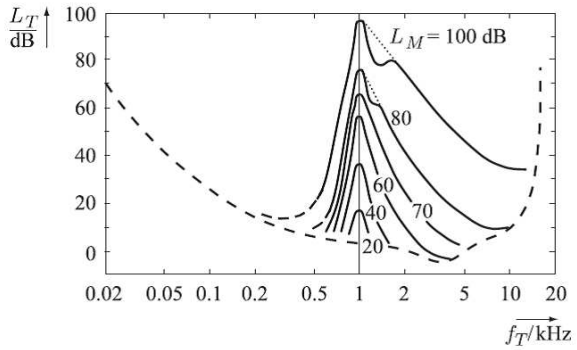


Abbildung 2.7: Maskierung [1, aus Kapitel 2.5.3]

Prinzipiell können all die Eigenschaften des menschlichen Gehörs, die beschrieben wurden, genutzt werden um das Rauschen in Sprachsignalen zu unterdrücken. Sie unterscheiden sich allerdings in der Komplexität ihrer Durchführung. So ist es sehr kom-

pliziert die genaue Frequenz-Ort-Transformation der Basilarmembran nachzubilden, wohingegen eine Implementierung der Maskierung einfacher ist. Daher wird in diesem Fall gezielt die Maskierung durch einen Algorithmus nachgebildet und genutzt, um das Rauschen auf natürliche Art und Weise in Sprachsignalen zu unterdrücken, ohne dass zusätzliche Störungen entstehen.

Kapitel 3

Störreduktion

Hörgeräteträger haben in Umgebungen, in denen viele verschiedene Geräusche gleichzeitig aktiv sind große Schwierigkeiten ihren gewünschten Gesprächspartner gut zu verstehen. Das liegt daran, dass die Geräusche alle gleichstark übertragen werden. In diesem Fall werden Filter benötigt, die das gewünschte Signal hervorheben, indem die störenden Quellen unterdrückt werden. Wie diese Filter aufgebaut sind und wie sie funktionieren wird in dem folgenden Kapitel erklärt.

3.1 Wiener Filter

Als Quelle diente für diesen Abschnitt [6, vgl. Kapitel 6]. Das Wiener Filter ist ein optimales Filter. Bei diesem wird eine Schätzung des Nutzanteils durch die Minimierung eines Fehlerkriteriums erreicht. Das Kriterium, das genutzt wird, ist der mittlere quadratische Fehler, *Mean Squared Error (MSE)*. In Abb. 3.1 ist ein LTI (Linear zeitin-

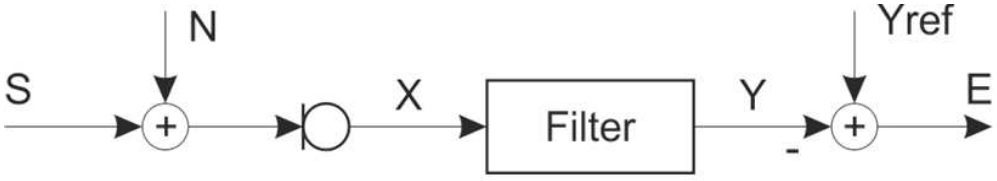


Abbildung 3.1: Schematische Übersicht des Wiener Filters

variantes) System dargestellt. Dabei wird ein klares Sprachsignal $s[k]$ mit dem Rauschsignal $n[k]$ überlagert. Diese beiden Signale zusammen bilden das Eingangssignal $x[k]$. Es wird angenommen, dass die beiden Signale $s[k]$ und $n[k]$ unkorreliert sind.

$$x[k] = n[k] + s[k] \tag{3.1}$$

Mittels der zeitdiskreten Fourier Transformation (DTFT) ($\mathcal{F}_*$) werden die Signale in den Frequenzbereich transformiert. Es ergibt sich mit $X(e^{j\Omega}) = \mathcal{F}_* \{x[k]\}$, $N(e^{j\Omega}) = \mathcal{F}_* \{n[k]\}$ und $S(e^{j\Omega}) = \mathcal{F}_* \{s[k]\}$ für die Gleichung des Eingangssignals in Abb. 3.1:

$$X(e^{j\Omega}) = N(e^{j\Omega}) + S(e^{j\Omega}) \tag{3.2}$$

Die Abb. 3.1 ist im Frequenzbereich dargestellt und auch die folgenden Betrachtungen und Erläuterungen werden im Frequenzbereich erfolgen. Für das Argument $e^{j\Omega}$ gilt: $\Omega = \frac{2\pi f}{f_A}$, mit f_A als Abtastfrequenz. Weiterhin entsprechen alle mit einem $\hat{(\cdot)}$ gekennzeichneten Symbole Schätzungen der tatsächlichen Werte. Das Ausgangssignal des Filters in Abb. 3.1 ist durch $Y(e^{j\Omega})$ dargestellt, das von dem gewünschten Signal $Y_{ref}(e^{j\Omega})$, in der Regel das klare Sprachsignal, abgezogen wird. Dadurch kann der Fehler $E(e^{j\Omega})$ berechnet werden.

$$E(e^{j\Omega}) = Y_{ref}(e^{j\Omega}) - Y(e^{j\Omega}) \tag{3.3}$$

Da das Ziel darin besteht, dass das Ausgangssignal dem gewünschten Signal entspricht, muss der berechnete Fehler $E(e^{j\Omega})$ minimiert werden. In den nächsten Schritten wird zunächst das Ausgangssignal $Y(e^{j\Omega})$ als Produkt aus der Übertragungsfunktion $H(e^{j\Omega})$ und dem Eingangssignal $X(e^{j\Omega})$ geschrieben. Das bedeutet:

$$Y(e^{j\Omega}) = H(e^{j\Omega}) \cdot X(e^{j\Omega}) \tag{3.4}$$

Wenn diese Bedingung in die Gleichung 3.3 eingesetzt wird ergibt sich für den Fehler:

$$E(e^{j\Omega}) = Y_{ref}(e^{j\Omega}) - H(e^{j\Omega}) \cdot X(e^{j\Omega}) \tag{3.5}$$

Das Ziel ist es nun eine Übertragungsfunktion $H(e^{j\Omega})$ zu berechnen, die den mittleren quadratischen Fehler $E(e^{j\Omega})$ minimiert. Dies führt zu:

$$\begin{aligned} E[|E(e^{j\Omega})|^2] &= E[Y_{ref}(e^{j\Omega}) - H(e^{j\Omega})X(e^{j\Omega})]^* \cdot [Y_{ref}(e^{j\Omega}) - H(e^{j\Omega})X(e^{j\Omega})] \\ &= E[|Y_{ref}(e^{j\Omega})|^2] - H(e^{j\Omega})E[Y^*_{ref}(e^{j\Omega})X(e^{j\Omega})] \\ &\quad - H^*(e^{j\Omega})E[X^*(e^{j\Omega})Y_{ref}(e^{j\Omega})] + |H(e^{j\Omega})|^2 E[|X(e^{j\Omega})|^2]. \end{aligned} \tag{3.6}$$

Mit den Annahmen, dass $\hat{S}_{xx}(e^{j\Omega}) = E[X(e^{j\Omega}) \cdot X(e^{j\Omega})]$ das Leistungsdichtespektrum von $X(e^{j\Omega})$ und $\hat{S}_{xy}(e^{j\Omega}) = E[X(e^{j\Omega})Y^*_{ref}(e^{j\Omega})]$ das Kreuzleistungsspektrum von $X(e^j\Omega)$ und $Y_{ref}(e^{j\Omega})$ ist, kann der quadratische Fehler nach [6, Kapitel 6.3] vereinfacht werden:

$$\begin{aligned} J_2 &= E[|E(e^{j\Omega})|^2] = \\ &= E[|Y_{ref}(e^{j\Omega})|^2] - H(e^{j\Omega})\hat{S}_{xy}(e^{j\Omega}) - H^*(e^{j\Omega})\hat{S}_{yx}(e^{j\Omega}) + |H(e^{j\Omega})|^2\hat{S}_{xx}(e^{j\Omega}) \end{aligned} \tag{3.7}$$

Um nun die optimale Übertragungsfunktion $H(e^{j\Omega})$ zu finden, die den Fehler minimiert, muss die Gleichung 3.7 nach $H(e^{j\Omega})$ abgeleitet und gleich null gesetzt werden. Wenn das Ergebnis direkt nach $H(e^{j\Omega})$ aufgelöst wird, erhält man die allgemeine Form des Wiener Filters im Frequenzbereich:

$$H(e^{j\Omega}) = \frac{\hat{S}_{yx}(e^{j\Omega})}{\hat{S}_{xx}(e^{j\Omega})} \tag{3.8}$$

Mit der Annahme, dass bei der Rauschunterdrückung das Referenzsignal $Y_{ref}(e^{j\Omega})$ dem gewünschten Signal, dem klaren Sprachsignal $S(e^{j\Omega})$, entspricht und das Gesamtsignal $X(e^{j\Omega})$ durch die Gleichung 3.2 ersetzt wird, ergibt sich für das Kreuzleistungsspektrum $\hat{S}_{yx} = \hat{S}^*_{xy} = E[(X(e^{j\Omega})Y^*_{ref}(e^{j\Omega}))^*]$:

$$\begin{aligned} \hat{S}_{yx} &= E[Y_{ref}(e^{j\Omega})X^*(e^{j\Omega})] \\ &= E[S(e^{j\Omega})(S(e^{j\Omega}) + N(e^{j\Omega}))^*] \\ &= E[S(e^{j\Omega})S^*(e^{j\Omega})] + E[S(e^{j\Omega})N^*(e^{j\Omega})] \\ &= E[S(e^{j\Omega})S^*(e^{j\Omega})] = \hat{S}_{ss}(e^{j\Omega}) \end{aligned} \tag{3.9}$$

Das Kreuzleistungsdichtespektrum $\hat{S}_{sn}(e^{j\Omega})$ fällt weg, da angenommen wird, dass die beiden Signale unkorreliert sind. Der nächste Schritt ist die Gleichung 3.2 ebenfalls durch ihre Leistungsspektren darzustellen.

$$\hat{S}_{xx}(e^{j\Omega}) = \hat{S}_{nn}(e^{j\Omega}) + \hat{S}_{ss}(e^{j\Omega}) \tag{3.10}$$

Als nächstes werden nun die beiden vorhergehenden Gleichungen in die Gleichung 3.8 eingesetzt. Daraus ergibt sich für die Übertragungsfunktion $H(e^{j\Omega})$:

$$H(e^{j\Omega}) = \frac{\hat{S}_{ss}(e^{j\Omega})}{\hat{S}_{nn}(e^{j\Omega}) + \hat{S}_{ss}(e^{j\Omega})} \tag{3.11}$$

$H(e^{j\Omega})$ ist reellwertig, nicht negativ und gerade, weil die Spektren $\hat{S}_{nn}(e^{j\Omega})$ und $\hat{S}_{ss}(e^{j\Omega})$ beide größer null sind und gerade Symmetrie besitzen. Es werden nur Werte zwischen 0 und 1 angenommen, wobei $H(e^{j\Omega}) \approx 0$ für niedrige SNR Werte gilt. Für sehr hohe SNR Werte gilt $H(e^{j\Omega}) \approx 1$ [6, Kapitel 6.5]. Wenn nun noch in die vorhergehende Gleichung 3.11 für $\hat{S}_{ss}(e^{j\Omega}) = \hat{S}_{xx}(e^{j\Omega}) - \hat{S}_{nn}(e^{j\Omega})$ eingesetzt wird kann die Formel für das Wiener Filter im Frequenzbereich umformuliert werden als:

$$\begin{aligned} H(e^{j\Omega}) &= \frac{\hat{S}_{xx}(e^{j\Omega}) - \hat{S}_{nn}(e^{j\Omega})}{\hat{S}_{xx}(e^{j\Omega})} \\ &= 1 - \frac{\hat{S}_{nn}(e^{j\Omega})}{\hat{S}_{xx}(e^{j\Omega})} \end{aligned} \tag{3.12}$$

3.2 Spektrale Subtraktion

Die Informationen für diesen Abschnitt basieren auf [6, Kapitel 5]. Bei der spektralen Subtraktion geht es darum, eine Schätzung des Sprachsignals zu erhalten, indem vom Gesamtsignal eine Schätzung des Rauschsignals subtrahiert wird. Die Phase wird dabei außer Acht gelassen, da das menschlichen Gehör unempfindlich gegenüber Phasenverschiebungen ist und diese sehr schwer zu schätzen ist, wodurch nur Leistungen voneinander abgezogen werden. Die Algorithmen der spektralen Subtraktion besitzen eine bekannte Vorgehensweise um das Hintergrundrauschen zu entfernen. Sie wird häufig gewählt, da es eine Methode ist, welche sehr einfach implementiert werden kann und

die durch Variationen der Parameter sehr flexibel ist. Als Ausgangsmodell wird ebenfalls wieder das Signalmodell aus Abb. 3.1 im vorhergehenden Abschnitt 3.1 verwendet. Daher lässt sich für das Mikrofonsignal folgende Gleichung aufstellen:

$$X(e^{j\Omega}) = S(e^{j\Omega}) + N(e^{j\Omega}) \tag{3.13}$$

Auch hier wird die Gleichung wieder im Leistungsdichtespektrum betrachtet, wobei auch hier das Symbol $\hat{(\cdot)}$ eine Schätzung darstellt. So ergibt sich für das verbesserte Sprachsignal:

$$\hat{S}_{ss}(e^{j\Omega}) = \hat{S}_{xx}(e^{j\Omega}) - \hat{S}_{nn}(e^{j\Omega}) \tag{3.14}$$

Diese Gleichung beschreibt den Algorithmus der spektralen Subtraktion. Sie kann aber auch durch eine Übertragungsfunktion nach [6, Kapitel 5.1] ausgedrückt werden,

$$\hat{S}_{ss}(e^{j\Omega}) = |\tilde{H}(e^{j\Omega})|^2 \hat{S}_{xx}(e^{j\Omega}) \tag{3.15}$$

wobei $\tilde{H}(e^{j\Omega})$ durch Einsetzen der Gleichung 3.15 in 3.14 berechnet werden kann. Es ergibt sich für die Übertragungsfunktion:

$$\begin{aligned} |\tilde{H}(e^{j\Omega})|^2 \hat{S}_{xx}(e^{j\Omega}) &= \hat{S}_{xx}(e^{j\Omega}) - \hat{S}_{nn}(e^{j\Omega}) \\ |\tilde{H}(e^{j\Omega})|^2 &= 1 - \frac{\hat{S}_{nn}(e^{j\Omega})}{\hat{S}_{xx}(e^{j\Omega})} \\ \tilde{H}(e^{j\Omega}) &= \sqrt{1 - \frac{\hat{S}_{nn}(e^{j\Omega})}{\hat{S}_{xx}(e^{j\Omega})}} \end{aligned} \tag{3.16}$$

Die Übertragungsfunktion nimmt nur Werte im Bereich von 0 bis 1 an und ist immer positiv. Negative Werte können durch fehlerhafte Schätzungen des Rauschens auftreten. Als letzter Punkt wird die Beziehung nach [1, Kapitel 11.3.2] zwischen den beiden vorgestellten Filtern erklärt. So kann man erkennen, dass die Übertragungsfunktion $\tilde{H}(e^{j\Omega})$ der spektralen Subtraktion die Wurzel aus der Übertragungsfunktion $H(e^{j\Omega})$ des Wiener Filters ist.

$$\tilde{H}(e^{j\Omega}) = \sqrt{1 - \frac{\hat{S}_{nn}(e^{j\Omega})}{\hat{S}_{xx}(e^{j\Omega})}} = \sqrt{H(e^{j\Omega})} \tag{3.17}$$

Wenn die spektrale Subtraktion in Verbindung mit dem Wiener Filter verwendet wird, ergibt sich nach [1, Kapitel 11.3.2] folgende Gleichung:

$$H(e^{j\Omega}) = \frac{\hat{S}_{xx}(e^{j\Omega}) - \hat{S}_{nn}(e^{j\Omega})}{\hat{S}_{xx}(e^{j\Omega})} = 1 - \frac{\hat{S}_{nn}(e^{j\Omega})}{\hat{S}_{xx}(e^{j\Omega})} \tag{3.18}$$

Diese und andere Variationen der spektralen Übertragungsfunktion werden in einer allgemeinen Übertragungsfunktion zusammengefasst:

$$\tilde{H}(e^{j\Omega}) = \left[1 - \left(\frac{\hat{S}_{nn}(e^{j\Omega})}{\hat{S}_{xx}(e^{j\Omega})}\right)^{\gamma_1}\right]^{\gamma_2} \tag{3.19}$$

Wählt man nun $\gamma_1 = 1$ und $\gamma_2 = 1$ so ergibt sich aus der allgemeinen Übertragungsfunktion die Funktion für den Wiener Filter, wenn $\gamma_1 = 1$ und $\gamma_2 = \frac{1}{2}$ gewählt wird ergibt sich die Übertragungsfunktion für die spektrale Subtraktion. Die allgemeine Übertragungsfunktion aus Gleichung 3.19 ist in der Realität schwierig zu implementieren, da bei den Berechnungen negative Werte auftreten können. Um diese zu verhindern wurde eine Übertragungsfunktion implementiert, die im Fall von negativen Werten einen Minimalwert annimmt.

$$\tilde{H}(e^{j\Omega}) = \max\left(\beta\left(\frac{\hat{S}_{nn}(e^{j\Omega})}{\hat{S}_{xx}(e^{j\Omega})}\right)^{\gamma_1}, 1 - \alpha\left(\frac{\hat{S}_{nn}(e^{j\Omega})}{\hat{S}_{xx}(e^{j\Omega})}\right)^{\gamma_1}\right)^{\gamma_2}, \tag{3.20}$$

Dabei stellt $\beta\left(\frac{\hat{S}_{nn}(e^{j\Omega})}{\hat{S}_{xx}(e^{j\Omega})}\right)^{\gamma_1}$ das Minimum und $1 - \alpha\left(\frac{\hat{S}_{nn}(e^{j\Omega})}{\hat{S}_{xx}(e^{j\Omega})}\right)^{\gamma_1}$ die parametrisierte Übertragungsfunktion des Störreduktionsfilters dar. Nachfolgend sollen die beiden Parameter α und β optimiert werden. Dies ist aber eine schwierige Aufgabe, da die beiden Parameter für verschiedene Eigenschaften im Signal verantwortlich sind. So wird α als *oversubtracting factor* bezeichnet, durch den das Kurzzeitspektrum des Signals stärker gedämpft wird gleichzeitig aber die Verzerrung des Nutzsignals zunimmt. Der zweite Parameter β wird als *spectral flooring* bezeichnet. Dieser verringert das Rauschen bei hohen Werten für β stark, wodurch nur noch ein niedriges Restrauschen vorhanden ist. Bei hohen Werten für β wird das Restrauschen sehr hoch sein. Allerdings führen kleine Werte zu erhöhten Artefakten, den sogenannten *musical tones*, sodass ein geeigneter Kompromiss zwischen den *musical tones* und der Störreduktion gefunden werden muss.

Dies verdeutlicht bereits die Schwierigkeit gute bzw. optimale Werte fÃ¼r die beiden Parameter zu finden. Deshalb versucht man die Parameter so zu wählen, dass das Spektrum des Restrauschens kleiner ist als der Masking Threshold des Gehörs. Somit wäre das Rauschen maskiert und für den Menschen unhörbar. Ein mögliches Verfahren zur Parameteroptimierung basierend auf dieser Idee wird nachfolgend genauer untersucht.

Kapitel 4

Verfahren zur Parameteroptimierung basierend auf menschlicher Wahrnehmung

Ein großer Nachteil der Rauschunterdrückung mit einkanaligen Filtern sind die sogenannten *musical tones*. Diese entstehen durch Fehler bei den Schätzungen und sind durch zufällige Peaks im Spektrum sichtbar. Eigenschaften dieser *musical tones* sind eine erhöhte Varianz und das Auftreten bei beliebigen Frequenzen. Auf diesem Gebiet wurden schon viele verschiedene Varianten getestet, aber es ist dennoch eine Verbesserung nötig, da es sehr schwierig ist das Rauschen zu unterdrücken, ohne dabei die Sprachverständlichkeit zu verschlechtern und ohne, dass dabei Sprachstörungen oder *musical tones* auftreten. Wenn es eine sehr verrauschte Situation ist, kann es durchaus passieren, dass das verbesserte Signal schlechter ist als das Originalsignal [7].

Das Ziel ist deshalb, die *musical tones* unhörbar zu machen. Dies soll durch die Nutzung des Wissens über das menschliche Gehör erreicht werden. Genauer gesagt wird in dieser Arbeit die Maskierung und der Mechanismus der *critical band Analyse* des Innenohrs, mit dem die Maskierung in Beziehung steht, genutzt. Durch die Berechnung eines Rauschmaskierungsthresholds werden diese Eigenschaften nachgebildet. Da der

Mensch ein additives Rauschen nicht wahrnimmt, solange es unterhalb der Threshold-grenze liegt ist das Ziel, die Parameter auf dem Threshold basierend so anzupassen, dass das Spektrum des Rauschens unterhalb des *Masking Thresholds* liegt. Dadurch soll erreicht werden, dass das Verhältnis zwischen Rauschreduzierung und den *musical tones* verbessert wird, sodass das Ergebnis für den Hörer angenehmer wird und das gewünschte Signal besser wahrnehmbar ist [7].

4.1 Berechnung des Masking Thresholds

Um diesen gewünschten *Masking Threshold* zu erhalten, wurde das Verfahren von [7] implementiert. Dieses impliziert vier verschiedene Teilschritte, die im Folgenden näher erklärt werden.

1. Frequenzbandanalyse: Der erste Schritt basiert auf dem Mechanismus des Innenohrs, der Einteilung der Basilarmembran in 24 Abschnitte, den sogenannten *bark scales* oder auch *critical bands* genannt, die schon im Abschnitt 2.2.2 beschrieben wurden. Hierzu wird für die Störreduktion die DFT (Diskrete Fourier Transformation) der Signale berechnet. Die Frequenzauflösung muss dabei hoch genug sein, um die Störung gut unterdrücken zu können. Die Berechnung der DFT liefert eine gleichmäßige Auflösung des Signals. Da aber die *critical bands* genutzt werden sollen, muss gezählt werden, wie viele Frequenzbins der gleichmäßigen Auflösung jeweils innerhalb eines *critical bands* liegen. Für die Einteilung der Basilarmembran wird die Tabelle 4.1 der *bark scales* nach [8] herangezogen. Anschließend werden in jedem *critical band* einzeln, entsprechend der zugehörigen Anzahl Frequenzbins, die Energien des DFT Spektrums, wie in [8] beschrieben, aufaddiert:

$$B(i) = \sum_{\omega=bl_i}^{bh_i} \hat{S}_{ss}(\omega) \tag{4.1}$$

$\hat{S}_{ss}(\omega)$ ist die Schätzung des Nutzsignals im Leistungsdichtespektrum. Der Wert bl_i beschreibt die untere Grenze des *critical bands* und bh_i entsprechend die obere Grenze

Band Nummer	Untere Kante	Mittlere Frequenz	Obere Kante
Hz	Hz	Hz	Hz
1	0	50	100
2	100	150	200
3	200	250	300
4	300	350	400
5	400	450	510
6	510	570	630
7	630	700	770
8	770	840	920
9	920	1000	1080
10	1080	1170	1270
11	1270	1370	1480
12	1480	1600	1720
13	1720	1850	2000
14	2000	2150	2320
15	2320	2500	2700
16	2700	2900	3150
17	3150	3400	3700
18	3700	4000	4400
19	4400	4800	5300
20	5300	5800	6400
21	6400	7000	7700
22	7700	8500	9500
23	9500	10500	12000
24	12000	14000	16000

Tabelle 4.1: Tabelle der *critical bands* [8]

des *critcal bands*. Da dies eine diskrete Form der kritischen Frequenzanalyse ist, wird die Maskierung nur von zwei Signalen, die im gleichen *critical band* liegen, berücksichtigt. Um aber ein kontinuierliches Bandspektrum zu erhalten, müssen diese Bereiche zwischen den *critical bands* berücksichtigt werden [8].

2. Streufunktion: Genau diese Bereiche zwischen den *critical bands* werden mit Hilfe einer Streufunktion $S(i)$ aus [9] berechnet:

$$S(i) = 10^{\frac{1}{10}(15.81+7.5(i+0.474)-17.5\sqrt{1+(i+0.474)^2})} \tag{4.2}$$

Der Buchstabe i steht hier für die Zahl des *critical bands*, das bedeutet die Funktion wird für jedes einzelne *critical band* berechnet. Diese Streufunktion $S(i)$ wird mittels einer Matrixmultiplikation mit $B(i)$ verrechnet. Dadurch können die Bereiche zwischen den *critical bands* ebenso berücksichtigt werden. Die Gleichung liefert [8]:

$$C(i) = B(i) * S(i) \tag{4.3}$$

$C(i)$ beschreibt die Spektrumsspreizung der *critical bands*.

3. Threshold Berechnung: Als nächstes wird der gesuchte Threshold berechnet. Dafür werden zusätzliche Berechnungen benötigt. Es wird zwischen zwei verschiedenen Thresholds unterschieden. Der tonhaltige Threshold, welcher laut [8] auf $(14.5 + i)$ dB unterhalb von $C(i)$ geschätzt wird, wobei i wieder der Nummer des *critical bands* entspricht, und der rauschhaltige Threshold, welcher nach [8] 5.5 dB unterhalb von $C(i)$ geschätzt wird. Um zu berechnen ob ein Signal eher tonhaltig oder rauschhaltig ist, wird die spektrale Flachheit (SFM: Spectral Flateness Measure) genutzt. Diese ist als Quotient aus dem geometrischem Mittelwert G_m und dem arithmetischem Mittelwert A_m definiert:

$$\mathrm{SFM_{dB}} = 10 \log 10 \frac{G_m}{A_m} \tag{4.4}$$

Mit Hilfe dieser Angabe kann nun die Tonhaltigkeit a des Signals bestimmt werden:

$$a = \min\left(\frac{\mathrm{SFM_{dB}}}{\mathrm{SFM_{dBMax}}}\right) \tag{4.5}$$

Der Wert von SFM_{dBMax} liegt bei -60 dB. Wenn SFM_{dB} diesen Wert annimmt, spricht man von einem kompletten Tonsignal, nimmt SFM_{dB} hingegen einen Wert von 0 dB an, lässt das darauf schließen, dass nur Rauschen vorliegt. So kann ein Offset $O(i)$ des Signals nach [8] pro *critical band* i berechnet werden, das durch den Parameter a die beiden Thresholdoffsets, mit $(14.5 + i)$dB für ein tonhaltiges Signal und mit 5.5 dB für ein rauschhaltiges Signal miteinander gewichtet:

$$O(i) = a(14.5 + i) + (1 - a)5.5 \tag{4.6}$$

Jetzt kann der eigentliche Masking Threshold berechnet werden. Auch hierfür wurde eine Funktion aus [8] verwendet:

$$T(i) = 10^{\log_{10}(C(i)) - (\frac{O(i)}{10})} \tag{4.7}$$

Dabei wird der Offset $O(i)$ von der Funktion $C(i)$ abgezogen um so den gewünschten Masking Threshold zu erhalten.

4. Threshold zurückrechnen: Als letzter Schritt muss der Threshold wieder auf die eigentliche Länge des Signals zurückgerechnet werden. Dazu ist es wichtig zu wissen, wie viele Frequenzbins in einem *critical band* liegen. Dann kann der Wert des Thresholds des jeweiligen *critical bands* der Anzahl der Freuquenzbins entsprechend oft nacheinander eingetragen und so wieder auf die eigentliche Länge des Signals zurückgerechnet werden. All diese Schritte werden normalerweise anhand des klaren Sprachsignals berechnet. Ist dieses nicht verfügbar, kann mittels einer spektralen Subtraktionsmethode eine Schätzung des klaren Signals berechnet werden.

4.2 Anpassung der Parameter

Das Wissen über die Anpassung der Parameter stammt aus [7] und [6, Kapitel 5.11]. Meistens werden die Parameter experimentell, basierend auf dem SNR (Signal-to-Noise-Ratio), oder mit dem *Minimal Mean Square Error* (MMSE) berechnet. Bei diesen beiden Methoden werden allerdings die Wahrnehmungseigenschaften des menschlichen

Gehörs nicht beachtet. Genau diese sollen aber genutzt werden um das Rauschen zu maskieren und zugleich die Verständlichkeit zu verbessern. Deshalb ist das Ziel die beiden Parameter α und β, aus der Gleichung 3.20, basierend auf dem Pegel des Maskierungsthresholds, anzupassen. Unter Berücksichtigung dieser Feststellungen wurde eine Funktion für die Parameter basierend auf dem normierten Threshold $T(e^{j\Omega})$ erstellt. Dabei gilt: ist der Masking Threshold hoch, wird das Eigenrauschen auf natürliche Art und Weise maskiert und es muss nichts getan werden um das Rauschen zu verringern, sodass die Parameter ihre Minimalwerte annehmen können. Ist der Threshold niedrig bedeutet das, dass das Rauschen störend ist und entfernt werden muss. Daher müssen die Parameter erhöht werden und ihre Maximalwerte annehmen. Um dies zu realisieren, müssen die beiden Parameter α und β eine Funktion des Thresholds sein. Diese Funktion ist in [6, Kapitel 5.11] zu finden:

$$F_\alpha(e^{j\Omega}) = \begin{cases} \alpha_{max} & \text{für } T(e^{j\Omega}) = T(e^{j\Omega})_{min} \\ \alpha_{min} & \text{für } T(e^{j\Omega}) = T(e^{j\Omega})_{max} \end{cases} \tag{4.8}$$

$$F_\beta(e^{j\Omega}) = \begin{cases} \beta_{max} & \text{für } T(e^{j\Omega}) = T(e^{j\Omega})_{min} \\ \beta_{min} & \text{für } T(e^{j\Omega}) = T(e^{j\Omega})_{max} \end{cases} \tag{4.9}$$

Die Berechnung der Parameter muss für jedes *critical band* oder eben jeden Frequenzbin i des normierten Signals erfolgen. Es muss daher für jedes *critical band* i geprüft werden, ob der, nach Gleichung 4.7 berechnete, Threshold dem Minimal- oder dem Maximalwert des Thresholds entspricht. Falls dies der Fall ist, werden die Parameter entsprechend auf ihre Maximal- oder Minimalwerte gesetzt. Ist dies nicht der Fall werden die Werte für die Parameter mittels einer Umrechnung vom größeren Zahlenbereich auf den kleineren ermittelt.

$$\alpha = \left((T(e^{j\Omega})_{max} - T(i)) \frac{\alpha_{max} - \alpha_{min}}{T(e^{j\Omega})_{max} - T(e^{j\Omega})_{min}} \right) + \alpha_{min} \tag{4.10}$$

$$\beta = \left((T(e^{j\Omega})_{max} - T(i)) \frac{\beta_{max} - \beta_{min}}{T(e^{j\Omega})_{max} - T(e^{j\Omega})_{min}} \right) + \beta_{min} \tag{4.11}$$

Der Wert $T(i)$ entspricht hierbei dem Threshold, nach Gleichung 4.7 berechnet, in dem entsprechenden Frequenzbin, der Wert $T(e^{j\Omega})$ entspricht den Minimal- beziehungsweise Maximalwerten des Thresholds, bezogen auf das gesamte Signal. Nachdem für jeden Frequenzbin nun die entsprechenden Parameter α und β berechnet wurden, können diese in die Gleichung 3.20 für dieÜbertragungsfunktion eingesetzt werden um somit die Ergebnisse für den Wiener Filter oder die spektrale Subtraktion zu erhalten. Da es sehr schwierig ist, einen geeigneten Trade-off zwischen der Störreduktion und den *musical tones*, für die beiden Parameter α und β, zu finden, wie schon in Abschnitt 3.2 beschrieben, ist es nicht möglich absolute Werte für die Grenzbereiche der beiden Parameter anzugeben. Diese müssen durch Tests und Auswertungen analysiert werden [6, Kapitel 5.11].

Kapitel 5

Evaluation und Auswertung

Zur Auswertung des Algorithmus wurden drei verschiedene Maße verwendet. Es wurden Signal-to-Interference-Ratio (SIR) Werte berechnet, die ein Verhältnis der Leistung des Nutzsignals zur Leistung des Rauschens liefern. Das bedeutet, wenn das Rauschen dominiert, also einen größeren Anteil hat als das Nutzsignal, ergibt sich für das entsprechende SIR Verhältnis ein kleiner Wert. Ist der Rauschanteil hingegen gering, und das Sprachsignal dominiert, werden für das Verhältnis große Werte erreicht. Es wurden zwei SIR Werte ermittelt: am Eingang, $\mathrm{SIR}_{\mathrm{in}}$, und am Ausgang, $\mathrm{SIR}_{\mathrm{out}}$, des Systems. Die genaue Berechnung der beiden Werte ist in den folgenden Gleichungen zu sehen:

$$\mathrm{SIR}_{\mathrm{in}} = 10 \log \left(\frac{E[s^2[k]]}{E[n^2[k]]} \right) \mathrm{dB} \tag{5.1}$$

$$\mathrm{SIR}_{\mathrm{out}} = 10 \log \left(\frac{E[s_{out}^2[k]]}{E[n_{out}^2[k]]} \right) \mathrm{dB} \tag{5.2}$$

In schwierigeren Szenarien, in denen ein zusätzliches Hintergrundrauschen vorhanden ist, wird anstelle des SIR ein sogenanntes SINR (Signal-to-Interference-Noise-Ratio) berechnet. Hierbei müssen alle vorhandenen Rauschsignale addiert werden um den Anteil des Rauschens zu berechnen. Für die Auswertung wurde aus den beiden Werten, $\mathrm{SIR}_{\mathrm{in}}$ und $\mathrm{SIR}_{\mathrm{out}}$, die Differenz gebildet. Da man davon ausgeht, dass am Eingang ein stark verrauschtes Signal anliegt, welches dann einen relativ kleinen SIR Wert besitzt,

und am Ausgang ein geringer verrauschtes Signal anliegt, sollte der SIR Wert folglich groß sein. Somit liefert die Differenz einen positiven, möglichst großen Wert, wenn das Rauschen vom System gut unterdrückt wurde. Es ist daher sozusagen ein Gewinn im SIR zu erkennen, weshalb der Wert der Differenz als $\mathrm{SIR}_{\mathrm{gain}}$ bezeichnet wird. Die genaue Berechnung des $\mathrm{SIR}_{\mathrm{gain}}$ ist in folgender Gleichung beschrieben:

$$\mathrm{SIR}_{\mathrm{gain}} = \mathrm{SIR}_{\mathrm{out}} - \mathrm{SIR}_{\mathrm{in}} \tag{5.3}$$

Ein weiteres Maß ist die Nutzsignalverzerrung (Speech Distortion SD). Diese gibt eine Leistung des Fehlers zwischen dem Nutzsignal $s[k]$ und der Schätzung des Nutzsignals $\hat{s}[k-\tau]$ normiert auf die Leistung des Nutzsignals an. Daraus ergibt sich für die Nutzsignalverzerrung:

$$\mathrm{SD} = 10\log\left(\frac{E[(s[k]-\hat{s}[k-\tau])^2]}{E[s^2[k]]}\right)\mathrm{dB}, \tag{5.4}$$

wobei τ zur Phasenkompensation des Signalverarbeitungssystems benötigt wird. Hier gilt: die Schätung des Nutzsignals $\hat{s}[k-\tau]$ soll möglichst gleich dem Nutzsignal $s[k]$ sein. Im Optimalfall, entspricht die Schätzung dem Nutzsignal und es ergibt sich für die Differenz im Zähler ein Wert von 0, also insgesamt für die Nutzsignalverzerrung ein Wert im Bereich von $-\infty$. Da dies aber nie der Fall sein wird, ist das Ziel eine Schätzung zu erhalten die dem Nutzsignal möglichst ähnlich ist. Daher sollte die Berechnung der Nutzsignalverzerrung sehr kleine Werte liefern. Als drittes Maß wurde das Perceptual Similarity Measure (PSM) aus [10] herangezogen. Es muss berücksichtigt werden, dass alle nachfolgenden Berechnungen unter der Annahme, dass alle Signale bekannt sind, durchgeführt wurden. Um ein umfangreiches Ergebnis zu erhalten wurden die Berechnungen in verschiedenen Räumen getestet. Das einfachste Szenario konnte durch den hallarmen Raum (HR), mit 50 ms Nachhallzeit, nachgebildet werden. Zur realitätsnahen Bewertung wurden auch noch weitere Räume zur Auswertung herangezogen. Dabei handelt es sich um typische wohnzimmerähnliche Umgebungen mit Nachhallzeiten von 200 ms und 400 ms. Des Weiteren wurde auch die Anzahl der Quellen variiert um verschiedene Szenarien zu testen. So wurde die Anzahl von an-

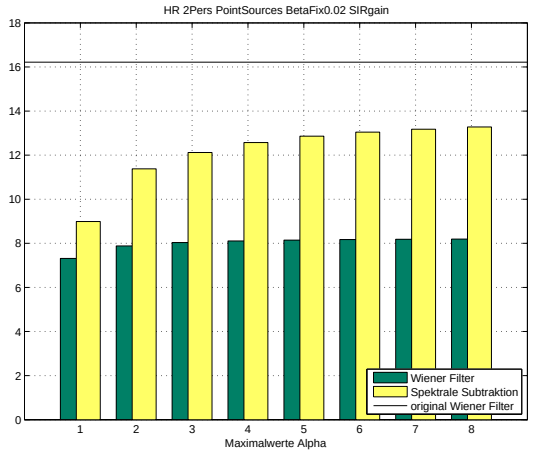
fangs zwei Quellen, eine Nutz- und eine Störquelle, auf insgesamt fünf Quellen, eine Nutz- und vier Störquellen, erhöht. Um sehr schwierige Szenarien zu realisieren, wurde ein zusätzliches Hintergrundrauschen (Babble Noise) hinzugefügt. Da das Ziel darin besteht, die besten Werte für die Parameter α und β zu finden wurden zunächst die Minimalwerte festgelegt. Hierbei wurde für $\alpha_{min} = 0.5$ gewählt, da es für Hörgeräteanwendungen wichtiger ist, möglichst verzerrungsfreie Signale zur Verfügung zu stellen anstatt eine große Rauschunterdrückung zu erhalten. Da aber für $\alpha > 1$ gilt, dass die Unterdrückung im Vordergrund steht, wurde α hier kleiner als 1 gewählt. Für den zweiten Parameter gilt: $\beta_{min} = 0$, da dieser als untere Grenze zum negativen dient, was auch in der implementierten Gleichung 3.20 zu erkennen ist. Daher wurde β_{min} so klein wie möglich gewählt. Zur Ermittlung der Maximalwerte der Parameter wurden die verschiedenen Szenarien nachgebildet. Hierbei wurde dann immer einmal der Parameter $\beta_{max} = 0.02$ als fester Wert angenommen und α_{max} wurde von 1 bis 8 variiert und ebenso wurde danach der Parameter $\alpha_{max} = 1$ als fester Wert angenommen und β_{max} von 0.01 bis 0.09 variiert. Die festen Werte der beiden Parameter wurden durch vorhergehende Tests ermittelt.

Es gilt für alle folgenden Szenarien, dass der Parameter β_{max} als fester Wert angenommen und der Parameter α_{max} variiert wird. Der Parameter α_{max} ist auf der x-Achse der Graphen angetragen. Weiterhin gilt, dass die Werte aller nachfolgenden Graphen immer für die spektrale Subtraktion (gelb) und für das modifizierte Wiener Filter (grün) in Relation zum normalen Wiener Filter, ohne Parameteranpassung, (blaue Linie) angegeben sind. Zu Beginn soll das einfachste Szenario beschrieben werden. Es gilt für dieses Szenario Tabelle 5.1. In der Abb. 5.1 (a) ist für den $\mathrm{SIR_{gain}}$ Wert gut zu

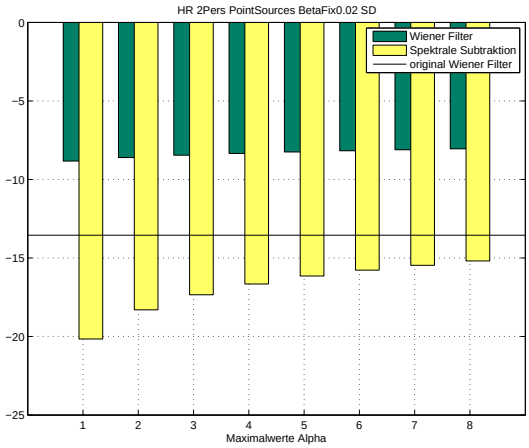
Nachhallzeit	50ms
Quellenanzahl	2
$\mathrm{SIR_{in}}$	0 dB
α_{min}	0.5
β_{min}	0
β_{max}	0.02

Tabelle 5.1: Parameter für Szenario mit Nachhallzeit 50 ms und 2 Punktquellen

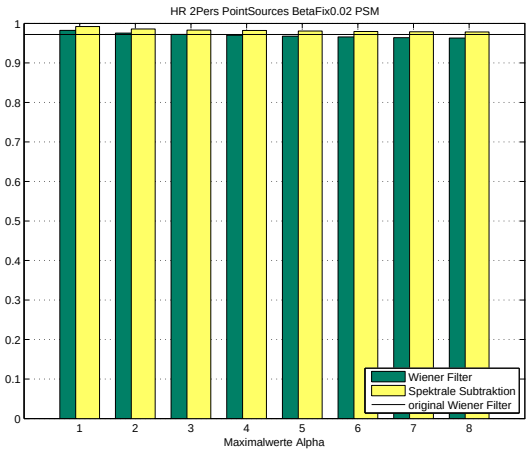
erkennen, dass sowohl das modifizierte Wiener Filter als auch die spektrale Subtraktion positive Werte liefern. Die Werte der spektralen Subtraktion sind allerdings im Vergleich zum modifizierten Wiener Filter wesentlich größer. Das bedeutet, es liegt in diesem Fall für die spektrale Subtraktion eine größere Rauschunterdrückung vor. Das Wiener Filter ohne Parameteranpassung hat eine noch größere Rauschunterdrückung als die spektrale Subtraktion, weshalb gefolgert werden könnte, dass das normale Wiener Filter besser funktioniert. Allerdings sollte auch die Nutzsignalverzerrung in Abb. 5.1 (b) dazu betrachtet werden, denn hier liefert die spektrale Subtraktion wesentlich bessere Werte als das normale Wiener Filter und auch das modifizierte Wiener Filter. Daher lässt sich insgesamt sagen, dass die spektrale Subtraktion, im Hinblick auf eine Hörgeräteanwendung, die besten Werte liefert, die auch in einem guten Wertebereich liegen. Diese Folgerung lässt sich auch erkennen, wenn Abb. 5.1 (c) betrachtet wird. Denn auch hier gilt, dass die Werte für die spektrale Subtraktion größer sind als für das modifizierte und nicht modifizierte Wiener Filter.



(a) SIR_{gain}



(b) SD



(c) PSM

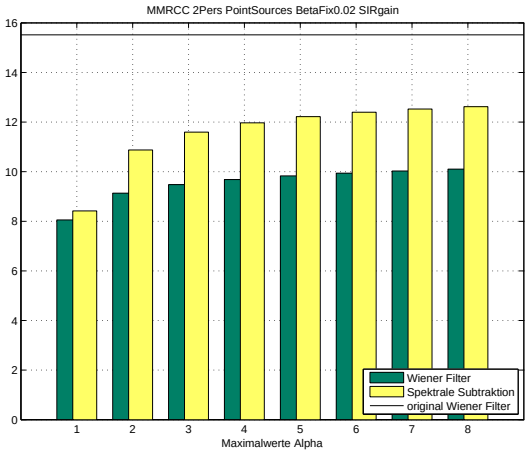
Abbildung 5.1: Nachhallzeit 50 ms - 2 Punktquellen

Als nächstes wird am Szenario nur der Raum geändert um zu verifizieren, was dieser für Auswirkungen hat. Es gelten daher die Werte aus der Tabelle 5.2. Auch in Abb. 5.2 (a) ist zu erkennen, dass das normale Wiener Filter den besten Wert für SIRgain liefert. Somit gilt wieder das normale Wiener Filter hat die beste Rauschunterdrückung. Allerdings sind auch die Werte der spektralen Subtraktion sehr gut und besser als die Werte des modifizierten Wiener Filters. Es darf aber wie schon im Szenario vorher die Nutzsignalverzerrung in Abb. 5.2 (b) nicht außer Acht gelassen werden, denn auch hier ist zu erkennen, das die spektrale Subtraktion bessere Werte liefert als das normale und das modifizierte Wiener Filter. Im Vergleich zum Szenario in Abb. 5.1 (a) und

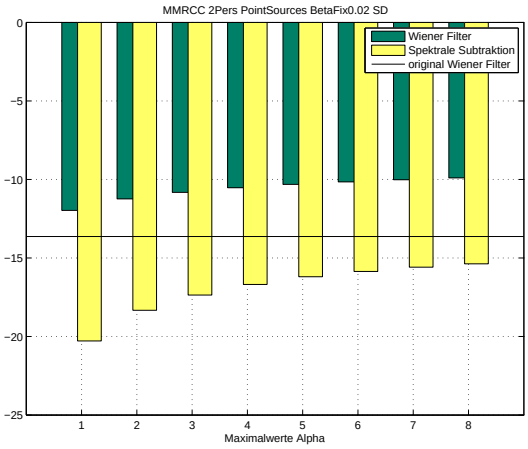
Nachhallzeit	200 ms
Quellenanzahl	2
SIR_{in}	0 dB
α_{min}	0.5
β_{min}	0
β_{max}	0.02

Tabelle 5.2: Parameter für Szenario mit Nachhallzeit 200 ms und 2 Punktquellen

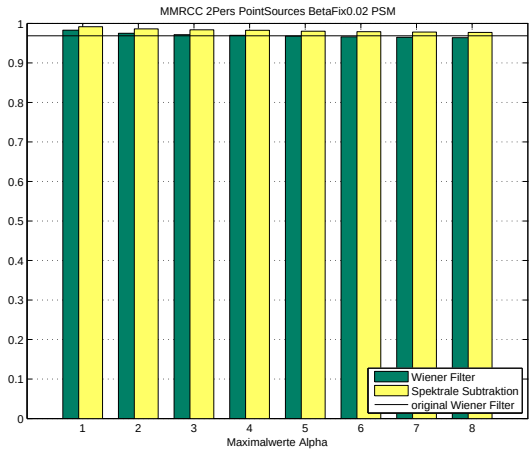
5.1 (b) gilt, die Werte des SIR_{gain} und der Nutzsignalverzerrung ändern sich für die spektrale Subtraktion und das normale Wiener Filter nur gering, wohingegen bei dem modifizierten Wiener Filter eine kleine Verbesserung der Werte zu erkennen ist. In der Abb. 5.2 (c) ist noch der Wert des PSM aufgetragen. Hier ist zu erkennen, dass ebenfalls die spektrale Subtraktion im Vergleich zu den beiden Wiener Filtern die größten Werte liefert.



(a) $\mathrm{SIR}_{\mathrm{gain}}$



(b) SD



(c) PSM

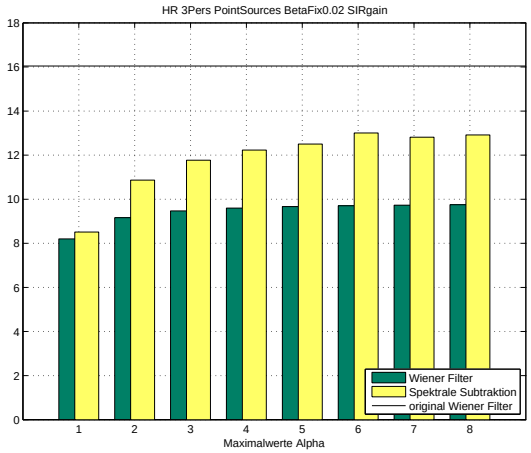
Abbildung 5.2: Nachhallzeit 200 ms - 2 Punktquellen

Im nächsten Szenario soll betrachtet werden wie sich die Werte verändern, wenn die Anzahl der Quellen erhöht wird. So gelten für das nächste Szenario die Werte aus Tabelle 5.3. In der Abb. 5.3 (a) ist wieder der $\mathrm{SIR}_{\mathrm{gain}}$ Wert abgebildet. Bei diesem ist zu erkennen, dass die Werte des normalen Wiener Filters und der spektralen Subtraktion insgesamt ein bisschen geringer sind im Vergleich zur Abb. 5.1 (a), das modifizierte Wiener Filter hat allerdings eine leicht steigende Tendenz. Dennoch liefert die spektrale Subtraktion gute Werte, die besser sind als das modifizierte Wiener Filter. Das normale Wiener Filter hat auch hier den größten Anteil der Rauschunterückung. Aber wie auch in den Abbildungen davor gilt: die Nutzsignalverzerrung muss mit betrachtet werden.

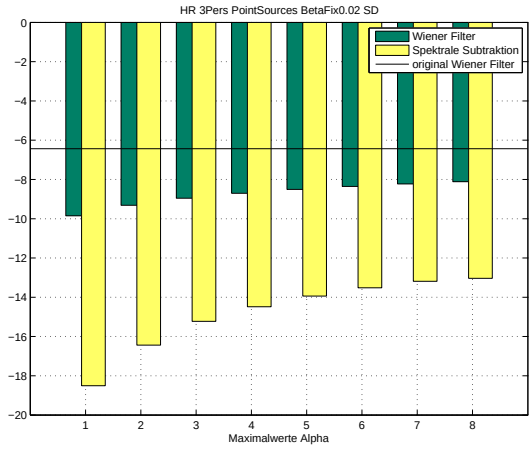
Nachhallzeit	50 ms
Quellenanzhal	3
SIR_{in}	-3 dB
α_{min}	0.5
β_{min}	0
β_{max}	0.02

Tabelle 5.3: Parameter für Szenario mit Nachhallzeit 50 ms und 3 Punktquellen

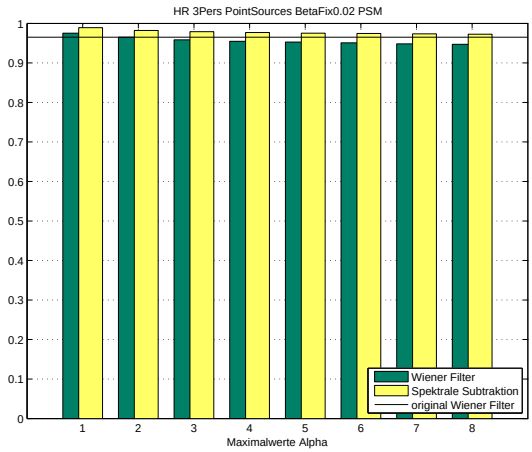
Diese ist in Abb. 5.3 (b) zu sehen. Dabei lässt sich erkennen, dass sich auch hier die Werte ein bisschen verschlechtert haben. Allerdings liegen die Werte für die spektrale Subtraktion immer noch in einem sehr guten Wertebereich und sind besser als die des modifizierten Wiener Filters. Der Wert des normalen Wiener Filters hingegen hat sich stark verschlechtert. In Abb. 5.3 (c) ist wiederum zu erkennen, dass auch hier die spektrale Subtraktion die besten Werte für die PSM liefert. So lässt sich insgesamt, auch im Vergleich zur Abb. 5.1, feststellen, dass sich die Werte des SIR_{gain} und der Nutzsignalverzerrung verschlechtert haben, die spektrale Subtraktion allerdings immer noch die besten Werte liefert.



(a) $\mathrm{SIR}_{\mathrm{gain}}$



(b) SD



(c) PSM

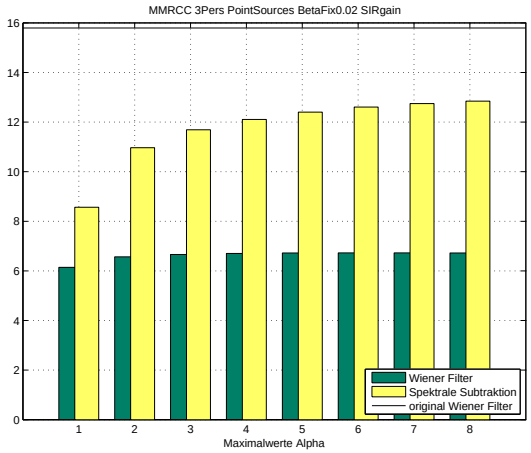
Abbildung 5.3: Nachhallzeit 50 ms - 3 Punktquellen

Auch hier soll nur der Raum geändert werden um mögliche Veränderungen festzustellen. Deshalb gelten für das folgende Szenario die Werte aus Tabelle 5.4. In Abb. 5.4 (a) ist der zugehörige $\mathrm{SIR}_{\mathrm{gain}}$ Wert abgebildet. Hier ist zu erkennen, dass die Werte für die spektrale Subtraktion und das normale Wiener Filter sich im Vergleich zur Abb. 5.3 (a) nur geringfügig ändern. Beim modifizierten Wiener Filter ist eine starke Abnahme zu erkennen. Daher gilt auch hier das normale Wiener Filter hat die größte Rauschunterdrückung. Die Rauschunterdrückung der spektralen Subtraktion ist allerdigns auch sehr gut. In Abb. 5.4 (b) ist zu erkennen, dass sich die Werte für das modifizierte Wiener Filter stark verschlechtert haben. Die spektrale Subtraktion hingegen hat nahezu

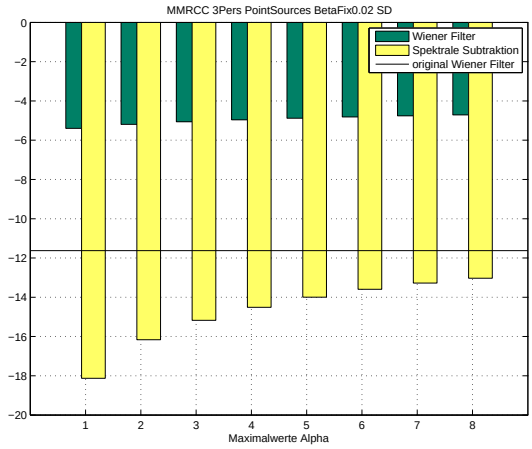
Nachhallzeit	200 ms
Quellenanzhal	3
SIR_{in}	-3 dB
α_{min}	0.5
β_{min}	0
β_{max}	0.02

Tabelle 5.4: Parameter für Szenario mit Nachhallzeit 200 ms und 3 Punktquellen

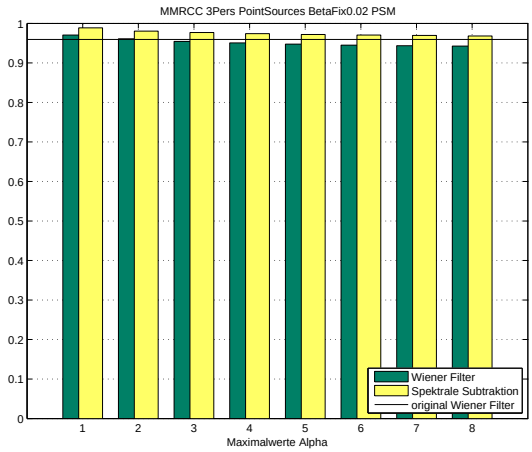
die gleichen Werte wie in Abb. 5.3 (b). Beim normalen Wiener Filter ist sogar eine Verbesserung der Werte zu erkennen. Da die Werte der spektralen Subtraktion aber immer noch besser sind als die des normalen Wiener Filters ist auch hier insgesamt die spektrale Subtraktion am besten. In der Abb. 5.4 (c) ist ebendies zu erkennen. Denn auch in diesem Fall gilt, die Werte der spektralen Subtraktion sind besser als die der beiden Wiener Filter.



(a) SIR_{gain}



(b) SD



(c) PSM

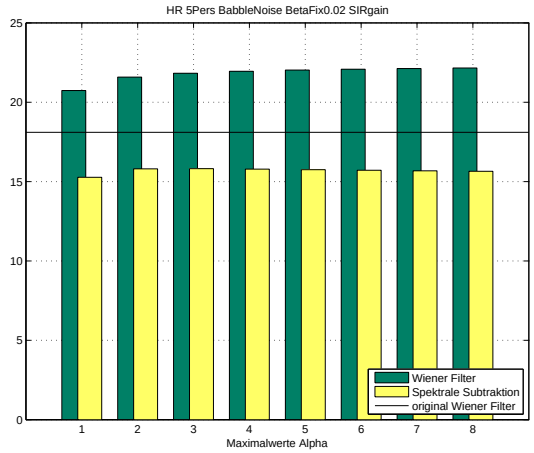
Abbildung 5.4: Nachhallzeit 200 ms - 3 Punktquellen

Als nächstes ist das schwierigste Szenario dargestellt. Es handelt sich dabei um die Maximalanzahl an Quellen mit Hinzunahme des zusätzliche Hintergrundrauschens. Es gelten daher für dieses die Werte aus Tabelle 5.5. Die zugehörigen Graphen sind in Abb. 5.5 dargestellt. In Abb. 5.5 (a) ist wiederum der SIR_{gain} Wert abgebildet. Im Vergleich zu den beiden ähnlichen Szenarien in Abb. 5.1 (a) und Abb. 5.3 (a), ist hier eine starke Veränderung zu sehen. So gilt zum ersten Mal, dass die größte Rauschunterdrückung durch das modofizierte Wiener Filter vorliegt. Dennoch ist zu erkennen, dass die Werte für die spektrale Subtraktion zwar im Vergleich am schlechtesten sind, sich aber im Vergleich zu den vorherigen Werten, aus Abb. 5.1 (a) und Abb. 5.3 (a), nicht

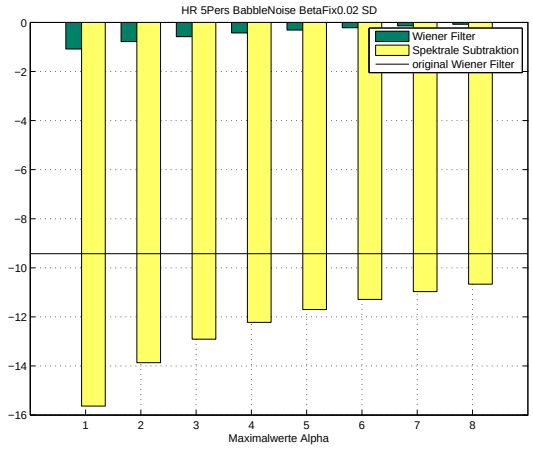
Nachhallzeit	50 ms
Quellenanzhal	5
Hintergrund	Babble Noise
SIR_{in}	-6 dB
α_{min}	0.5
β_{min}	0
β_{max}	0.02

Tabelle 5.5: Parameter für Szenario mit Nachhallzeit 50 ms und 5 Punktquellenuellen

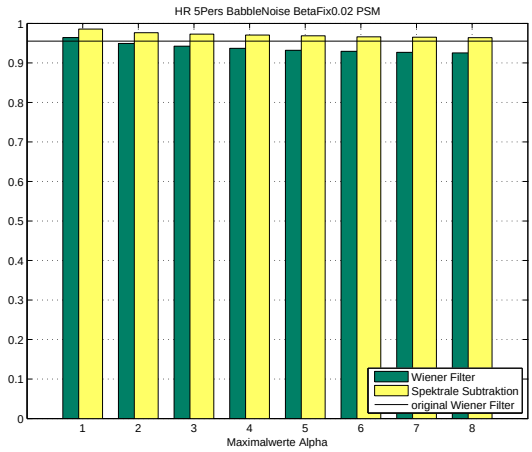
stark verändert haben. Außerdem gilt bei der Nutzsignalverzerrung, dass die Werte für das modifizierte Wiener Filter sehr schlecht sind. Daraus lässt sich erkennen, dass die Parameter für jedes Szenario für das Wiener Filter neu angepasst werden müssen um gute Ergebnisse zu erzielen. Für die spektrale Subtraktion hingegen zeigt sich nur eine geringfügige Verändeung der Werte. Diese leigen nach wie vor in einem guten Bereich. Das normale Wiener Filter besitzt auch gute Werte, aber dennoch ist der Unterschied bei der Nutzsignalverzerrung nicht zu verachten. Es ergibt sich dadurch wieder, dass die spektrale Subtraktion auch in diesem Szenario die besten Werte, im Hinblick auf eine Hörgeräteanwendung, liefert. Auch die PSM Werte sprechen für die spektrale Subtraktion, da diese, im Vergleich zu den beiden Wiener Filtern, die größeren Werte hat.



(a) SIR_{gain}



(b) SD



(c) PSM

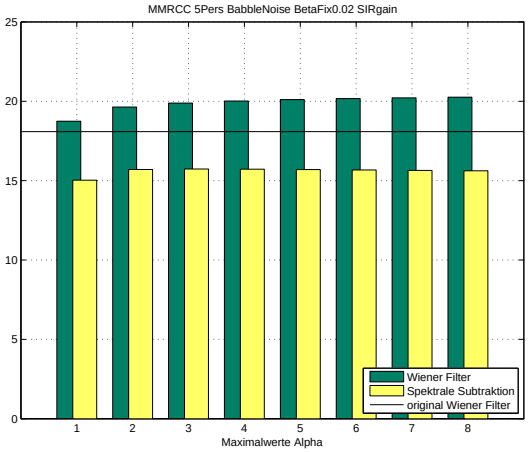
Abbildung 5.5: Nachhallzeit 50 ms - 5 Punktquellen mit zusätzlichem Hintergrundrauschen

Auch für dieses Szenario wird der hallarme Raum mit einem Raum mit 200 ms Verzögerung verglichen. Es gelten für das letzte Szenario die Werte aus Tabelle 5.6. Diese Ergebnisse sind in Abb. 5.6 dargestellt. Auch hier gilt für den SIR_{gain} Wert in Abb. 5.6 (a), dass das modifizierte Wiener Filter den größten Wert liefert. Wenn auch die Werte insgesamt geringer sind als in der Abb. 5.5 (a), so ist doch das Verhältnis des modifizierten Wiener Filters zu dem normalen und der spektralen Subtraktion annähernd gleich. Die Werte der Nutzsignalverzerrung sind allerdings im Vergleich zur Abb. 5.5 (b) wieder verbessert, besonders die des modifizierten Wiener Filters. Bei der spek-

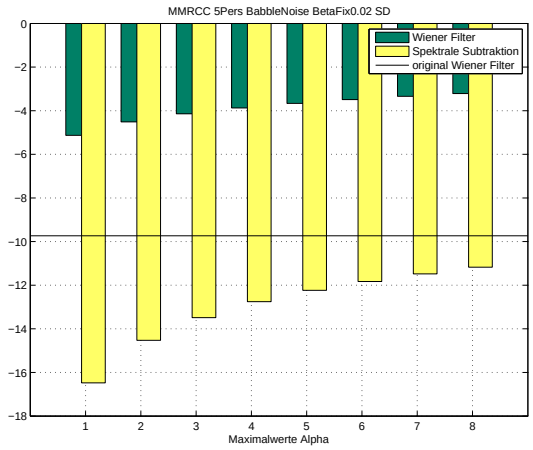
Nachhallzeit	200 ms
Quellenanzhal	5
Hintergrund	Babble Noise
SIR_{in}	-6 dB
α_{min}	0.5
β_{min}	0
β_{max}	0.02

Tabelle 5.6: Parameter für Szenario mit Nachhallzeit 200 ms und 5 Punktquellen

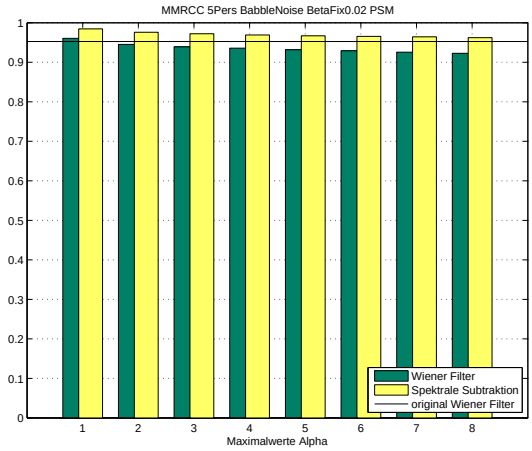
tralen Subtraktion ist keine große Veränderung zu erkennen, aber die Werte liegen noch immer in einem guten Bereich. Auch für dieses Szenario liefert die spektrale Subtraktion für die Werte des SIR_{gain} und die Nutzsignalverzerrung die besten Ergebnisse im Hinblick auf eine Hörgeräteanwendung. Auch in Abb. 5.6 (c) sind die Werte der spektralen Subtraktion im Vergleich zu den beiden Wiener Filtern besser.
Insgesamt lässt sich erkennen, dass die Werte für das Wiener Filter mit Parameteranpassung in den schwierigeren Szenarien immer schlechter werden. Deshalb müssen die Parameter für jedes Szenario erneut angepasst und geprüft werden um gute Ergebnisse zu erhalten. Die Ergebnisse für die spektrale Subtraktion sind im Vergleich zum normalen Wiener Filter, aber auch im Vergleich zum modifizierten Wiener Filter, wesentlich besser. Das kann zum Einen an dem direkten Wertevergleich abgelesen werden zum Anderen lässt sich auch bei einer Betrachtung der Gesamtheit feststellen, dass die Werte der spektralen Subtraktion nahezu immer im gleichen Wertebereich liegen. Dabei ist es egal welcher Raum es ist, wie viele Quellen genutzt werden, oder ob das zusätzliche Hintergrundrauschen aktiv ist oder nicht. Die Werte der spektralen Subtraktion ändern sich nur gering. Das modifizierte Wiener Filter weist bei einer Veränderung größere Abweichungen auf. Daraus kann der Schluss gezogen werden, dass die spektrale Subtraktion robuster und stabiler ist. So sollte für weitere Betrachtungen die spektrale Subtraktion verwendet werden.



(a) $\mathrm{SIR}_{\mathrm{gain}}$



(b) SD



(c) PSM

Abbildung 5.6: Nachhallzeit 200 ms - 5 Punktquellen mit zusätzlichem Hintergrundrauschen

Kapitel 6

Zusammenfassung

Die Grundidee dieser Arbeit war Eigenschaften des menschlichen Gehörs zu nutzen um eine bessere Störreduktion für Hörgeräteanwendungen zu erhalten. Dazu wurden verschiedene Schritte realisiert. Es wurde ein Threshold implementiert, der die Eigenschaften der Maskierung berücksichtigt, indem eine Relation der FFT Frequenzbins zu den *critical bands* des Innenohr hergestellt wurde. Weiterhin ist eine allgemeine Übertragungsfunktion, die durch Änderung der Parameter γ_1 und γ_2 sowohl das Wiener Filter als auch die spektrale Subtraktion realisiert, implementiert worden. Diese Übertragungsfunktion ist von den beiden Parametern α und β abhängig. Um die Maskierung in der Übertragungsfunktion zu berücksichtigen wurden die beiden Parameter als eine Funktion des Thresholds optimiert. In der Auswertung lieferte die spektrale Subtraktion die besten Ergebnisse für Hörgeräteanwendungen, da die Werte des Wiener Filters mit Parameteranpassung für schwierige Szenarien stark von der Wahl der Parameter abhängen und daher die Ergebnisse nicht so gut sind. Dies spricht für eine Verbesserung der Störreduktion, durch die Nutzung der psychoakustischen Eigenschaften, da die Werte der spektralen Subtraktion für alle Szenarien sehr stabil und robusst sind. Daher ist es naheliegend weitere Auswertungen und Tests dieses Verfahrens mit der spektralen Subtraktion durchzuführen. Es muss aber beachtet werden, dass es in dieser Arbeit lediglich darum ging, zu testen ob eine Nutzung der Eigenschaften des menschlichen Gehörs erfolgreich ist. Daher wurde in dieser Arbeit von dem Idealfall

ausgegangen, dass alle Signale, das gewünschte Sprachsignal sowie das Rauschsignal, bekannt sind. Auf dieser Annahme wurden auch alle Berechnungen durchgeführt. Für weitere Auswertungen müssen daher auch Schätzungen dieser Signale betrachtet werden um reale Ergebnisse dieses Verfahrens zu erhalten.

Literaturverzeichnis

[1] M. R. Vary, P., *Digital Speech Transmission*, 2005.

[2] (retrieved 15.08.2013) www.ein-klang-raum.de/img/ohr.jpg.

[3] (retrieved 21.08.2013) Hörschnecke. [Online]. Available: http://commons.wikimedia.org/wiki/File:Cochlea-crosssection-de.png

[4] R. Guski. (retrieved 29.08.2013) Wahrnehmen - ein lehrbuch (1996). [Online]. Available: http://eco.psy.ruhr-uni-bochum.de/ecopsy/download/Guski-Lehrbuch/Kap_4_2.html

[5] M. Gekle, *Physiologie*, 2010.

[6] P. C. Loizou, *Speech Enhancement*, 2007.

[7] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," *IEEE Transactions on Speech and Audio Processing*, pp. 126–137, March 1999.

[8] J. D. Johnston, "Transform coding of audio signals using perceptual noise criteria," *IEEE Journal on selected areas in communications*, pp. 314–323, February 1988.

[9] M. Schroeder, B. Atal, and J. Hall, "Optimizing digital speech coders by exploiting masking properties of the human ear," *Acoustical Society of America*, pp. 1647–1652, December 1979.

[10] K. B. Huber, R., "Pemo-q-a new method for objective audio quality assessment using a model of auditory perception," *IEEE Transaction on Audio, Speech, and Language Processing*, pp. 1902–1911, November 2006.