Universität Erlangen-Nürnberg
Laboratorium Nachrichtentechnik
Lehrstuhl für Multimediakommunikation und Signalverarbeitung

Studienarbeit

Untersuchungen eines neuen Ansatzes zur blinden Enthallung von Sprachsignalen

Holger Kunze

Betreuer Dipl.-Ing. Herbert Buchner
Prof. Dr.-Ing Walter Kellermann

Beginn: 1.4.03
Ende: 13.7.03

Ich versichere, dass ich die 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.

Bubenreuth, 13.7.03

Holger Kunze
Gartenfeld 8a
91088 Bubenreuth

Inhaltsverzeichnis

Zusammenfassung

Automatische Spracherkennungssysteme stellen eine immer wichtiger werdende Schnittstelle zwischen Mensch und Maschine dar. Während man mit Nahbesprechungsmikrofonen bereits sehr gute Worterkennungsraten erzielt, führen verschiedenartige Störungen bei freier Spracheingabe zu erheblichen Performanceverlusten. Eine dieser Störungen ist der Nachhall. Besonders der lange Nachhall, der z.B. in Büroräumen auftritt, lässt die Worterkennungsrate drastisch sinken.

Ein Standardansatz, dieses Problem zu lösen, ist die Verwendung eines Delay-and-Sum-Beamformers, durch den allerdings nur mäßige Erfolge erzielt werden. Wesentlich bessere Ergebnisse erhält man mit Hilfe der cepstralen Subtraktion, die wegen der großen Fensterlängen jedoch Probleme bei veränderlicher Umgebung hat.

In dieser Arbeit wird die Möglichkeit untersucht, ob und in welchem Umfang mit adaptiven Verfahren unter Verwendung von Statistik zweiter Ordnung eine blinde Enthallung möglich ist.

Als Ausgangspunkt für die Untersuchungen dient ein Verfahren, welches von Gillespie und Atlas entwickelt wurde. Dieses erwies sich jedoch aufgrund von Instabilität als ungeeignet für die Enthallung.

Dagegen kann durch eine Modifikation des Verfahrens der blinden Quellentrennung der durch Hall verursachte Verlust bei der Worterkennungsrate fast genauso reduziert werden, wie mit Hilfe der cepstralen Subtraktion.

Eine effiziente Möglichkeit dieses Verfahren zu implementieren stellt ein prädiktions-basierte Ansatz dar, der ebenfalls in dieser Arbeit untersucht wird.

1 Verwendete Formelzeichen

t	kontinuierliche Zeitvariable
k	diskreter Zeitindex
κ	diskrete Zeitverschiebung
l	Nummer des Mikrofonsignals
m	Zählindex für Fenster
$x_l(t), x_l[k]$	Mikrofonsignal
$\mathbf{x}[k]$	Vektor mit den letzten n Mikrofonsignalwerten beim Zeitindex k
$X(j\omega)$	Fouriertransformierte des Mikrofonsignals
X	Cepstrum des Mikrofonsignals
$s(t), s[k]$	Quellsignal
$S(j\omega)$	Fouriertransformierte des Quellsignals
S	Cepstrum des Quellsignals
$h_l(t), h_l[k]$	Raumimpulsantwort
$H(j\omega)$	Fouriertransformierte der Raumimpulsantwort
H	Cepstrum der Raumimpulsantwort
$w_l(t), w_l[k]$	Rekonstruktionsfilter für das l-te Mikrofon
$\mathbf{w}$	Vektordarstellung des Rekonstruktionsfilters
$\mathbf{W}$	Sylvestermatrixdarstellung des Rekonstruktionsfilters
$R_{yy}[\kappa]$	Autokorrelierte des Ausgangs
$\mathbf{R_{yy}}(m)$	Autokorrelationsmatrix des Ausgangssignals im m-ten Fenster
$R_{xly}[\kappa]$	Kreuzkorrelierte des l-ten Mikrofonsignals mit Ausgangssignal
$\mathbf{R_{xly}}(m)$	Kreuzkorrelationsmatrix des l-ten Mikrofonsignals mit Ausgangssignal im m-ten Fenster
$R_{dd}[\kappa]$	Wunschautokorrelierte des Ausgangssignals
ε	Fehlermaß
μ	Schrittweite bei der Adaption
$\beta(i,m)$	Gewichtsfunktion bei Ermittlung des Fehlermaßes im m-ten Fenster

2 Abkürzungen

ASR	Automatic Speech Recognition
BSS	Blind Source Separation
FFT	Fast Fourier Transform, schnelle Implementierung der Diskreten Fouriertransformation
LMS	Least Mean Square
NLMS	Normalized Least Mean Square
MFC	Mel Frequency Cepstrum
WER	Worterkennungsrate

3 Was ist Hall?

In vergangenen Jahren hat die automatische Spracherkennung große Fortschritte erfahren. Es wurden Diktiersysteme entwickelt, die zuverlässig mehrere zehntausend Wörter erkennen. Die Auskunft der Deutschen Bahn wurde auf Automaten umgestellt, die die Fragen der Anrufer beantworten, Operationstische in Kliniken können über Spracheingabe gesteuert werden.

Die Spracheingabe bei diesen Systemen muss in all diesen Fällen über Nahbesprechungsmikrofone (z.B. Headset) erfolgen, damit nur sehr wenige Störgeräusche aufgenommen werden. Die Störungen können additive Ursachen haben, wie z.B. andere Sprecher bzw. vorbeifahrende Autos, oder faltungsbedingt sein. Zur letzten Gruppe gehört der Hall.

Während bereits große Fortschritte bei der Reduzierung von Störgeräuschen durch weitere Sprecher erzielt worden sind, stellt Hall immer noch ein unbefriedigend gelöstes Problem dar. Deshalb sollen in dieser Arbeit verschiedene Ansätze zur Reduzierung von Hall untersucht werden.

Hall ist die Überlagerung eines Signals mit verschieden verzögerten und gedämpften Echos seiner selbst. In Formeln kann Hall als Faltung des Signals s(t) mit einer Raumimpulsantwort h(t) beschrieben werden

$$x(t) = s(t) * h(t) \quad , \tag{1}$$

wobei die Form der Raumimpulsantwort vom Raum abhängig ist.

Befindet man sich in einem großen Raum mit Betonwänden, an denen Schall gut reflektiert wird, erreichen viele starke Echos das Mikrofon. Zudem ist der Nachhall lang andauernd. So kann in großen Kirchen der Schall mehrere Sekunden lang wahrgenommen werden. In Büroräumen sind es immerhin noch einige hundert Millisekunden. Befindet man sich dagegen in hallarmen Räumen mit schlecht reflektierenden Wänden, beträgt die Nachhallzeit nur wenige zehn Millisekunden, in denen nur einige wenige schwache Echos vom Mikrofon aufgenommen werden. Gemein ist diesen Raumimpulsantworten jedoch, dass zunächst einige einzelne Echos auftreten, deren Zahl mit der Zeit ansteigt und schließlich rauschartigen Charakter besitzt, die Intensität jedoch exponentiell abnimmt.

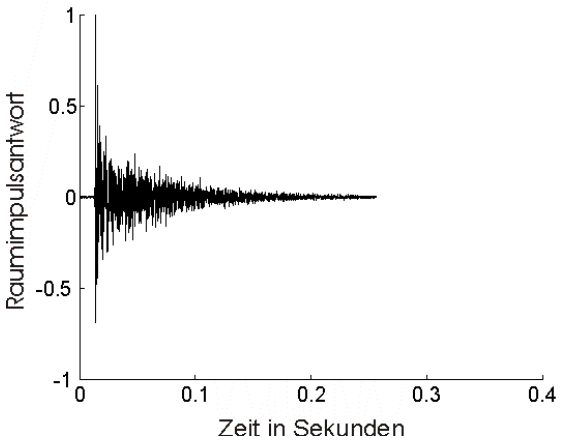

Abbildung 1: Raumimpulsantwort eines Büroraumes (2m x 3,5m x 3,5m), Abstand Quelle - Mikrofon 1m

Betrachtet man zwei Raumimpulsantworten, die im gleichen Raum mit verschieden positionierten Mikrofonen und gleicher Signalquelle aufgenommen worden sind, stellt man fest, dass sie sich im Allgemeinen nicht nur in der Grundverzögerung, sprich dem ersten Impuls, unterscheiden, sondern auch die Echos zu komplett verschiedenen Zeiten eintreffen und die rauschähnlichen Anteile unkorreliert sind. Die verschiedenen Grundlaufzeiten entstehen durch die verschiedenen Abstände der Quelle zu den Mikrofonen, so dass bei näheren Mikrofonen das Signal eher eintrifft als bei weiter entfernt stehenden. Die Eigenschaft, dass die Raumimpulsantworten in weiten Bereichen unkorreliert sind, lässt sich damit erklären, dass der Schall bereits dann verschiedene Wege nimmt, wenn die Mikrofone nur wenige Zentimeter voneinander entfernt stehen.

Wie bereits erwähnt, besitzen Hall und Nutzsignal die gleiche Quelle. Dadurch besitzen Hall und Nutzsignal ähnliche Eigenschaften. Deshalb ist es äußerst schwierig, die beiden Anteile zu trennen. Während bei additivem weißen Rauschen das Signal- zu Störleistungverhältnis mit Bandpassfiltern deutlich verbessert werden kann, erreicht man bei Hall durch diese Maßnahme nur geringe Verbesserungen.

Auch ist es nicht möglich, die Lösungen, die zur Verhinderung akustischer Rückkopplung entwickelt worden sind, bei dem vorliegenden Problem direkt anzuwenden, da hier im Gegensatz zu jenem Problem sowohl die Raumimpulsantwort als auch der ursprüngliche Signalverlauf nicht bekannt sind und beide durch das zu entwickelnde Verfahren zu schätzen sind.

Ebenfalls als untauglich für das Trennen von Nutzsignal und Hall erweisen sich die Verfahren zur blinden Entfaltung aus der Nachrichtenübertragung, da bei diesen zumindest der prinzipielle Signalverlauf bzw. der Grundimpuls bekannt sind, was die Entzerrung erleichtert.

Es muss also für das Problem der Enthallung von Sprache eine eigene Lösung gefunden werden. Erleichternd in dem vorliegenden Fall ist aber die Tatsache, dass der

Signalverlauf nicht genau wiederhergestellt werden muss. So kann man durch Hörversuche zeigen, dass Signale durch Allpassfilter “verzerrt” werden können, ohne dass das menschliche Ohr einen Unterschied hört. Ähnliches gilt für den automatischen Spracherkenner, der z.B. verschiedene Grundfrequenzen der Sprache ignoriert, wodurch es unter anderem möglich ist, dass verschiedene Menschen den gleichen Spracherkenner benutzen. Diese beiden Beispiele sollen verdeutlichen, dass der exakte Signalverlauf nur eine untergeordnete Rolle spielt und gefundene Lösungsvorschläge anhand von Hörversuchen und mit Spracherkennungssystemen beurteilt werden müssen.

4 Bewertungskriterien

Bevor der Hall und die Lösungsvorschläge zur Reduzierung von Hall in Mikrofonsignalen näher untersucht werden, sollen im Folgenden kurz die verwendeten Maße zur Beurteilung von Hall und Gütemaße zur Abschätzung der Fähigkeit der Algorithmen vorgestellt werden.

4.1 Höreindruck

Das erste Kriterium ist der Höreindruck, der einen ersten Anhaltspunkt für die Qualität des Algorithmus liefert. So hat sich bei Versuchen immer wieder erwiesen, dass nur schlechte Spracherkennungsperformance erreicht wurde, wenn schon beim Hören zusätzliche Artefakte oder Verzerrungen festgestellt wurden. Zudem sollte mit den Hörversuchen überprüft werden, ob sich der Algorithmus auch für die Enthallung von Signalen eignet, die später von Menschen wieder angehört werden sollen.

Die Beurteilung des Höreindrucks wurde während dieser Arbeit äußerst subjektiv ermittelt. Es wurde keine Befragung mehrerer Personen durchgeführt. Deswegen war die Bewertung des Algorithmus mit diesem Kriterium höchstens ein Anhaltspunkt und musste durch weitere Kriterien überprüft werden.

4.2 Systemabstand

Ein weiteres Maß ist der Systemabstand. Abweichend von der Definition bei der Kompensation akustischer Echos wird in dieser Arbeit der Systemabstand als die quadratische Norm der Impulsantwort definiert, die zwischen Signalquelle und Systemausgang gemessen wird. Der Systemabstand ist ein Maß für die Abweichung der Impulsantwort von einem verschobenen Einheitsimpuls, der das Ergebnis einer vollständigen Enthallung ist.

Die Impulsantwort wird zur Berechnung des Systemabstandes so normiert, dass das Maximum ihrer Absolutwerte eins ist. Auf diese Weise wird verhindert, dass verschiedene Dämpfungen der Raumimpulsantworten und der entwickelten Systeme, die mit einer automatischen Lautstärkeregelung ausgeglichen werden können, Einfluß auf den Systemabstand nehmen können.

$$A = \sum_{k=0}^{\infty} \left(\frac{h[k]}{\max(|h|)} \right)^2 \qquad (2)$$

Ein Einheitsimpuls hat somit den Systemabstand 1,0. Dies ist zugleich der kleinste Wert, da alle anderen Impulsantworten wegen der Normierung mindestens einen Peak der Größe 1 und dazu i.a. weitere Anteile aufweisen.

Dem Autor ist bewusst, dass dieses Kriterium allein nicht zur Bewertung der Algorithmen ausreicht, da es insbesondere nichts über die zeitliche Verteilung der Echos bzw. auch nichts über den Frequenzgang des Enthallungsfilters aussagt, der sich in den Versuchen später jedoch als ebenso wichtig herausstellte.

4.3 Nachhallzeit

Ein in der Raumakustik oft benutztes Maß ist die Nachhallzeit T_{60}. Die Nachhallzeit T_{60} gibt an, wie lange es dauert, bis die Schallenergie in einem Raum auf -60 dB des ursprünglichen Wertes abgeklungen ist.

Sie lässt sich mit Hilfe von Schröders Rückwärtsintegral berechnen:

$$r(t) = -10\log_{10}\left(\frac{\int_{t}^{\infty} h(\tau)^2 d\tau}{\int_{0}^{\infty} h(\tau)^2 d\tau}\right) \qquad (3)$$

Die Nachhallzeit T_{60} ist derjenige Wert, bei dem der Wert der Funktion r(t) 60 übersteigt. Da in den Simulationen meist abgeschnittene Raumimpulsantworten benutzt werden, kommt es bei der Bestimmung zu Ungenauigkeiten, die man durch Approximation der Funktion r(t) mittels einer linearen Funktion auszugleichen versucht.

4.4 Worterkennungsrate

Das letzte in dieser Arbeit verwendete Maß stellt die Worterkennungsrate (WER) dar. Erst mit ihr kann eine endgültige Aussage über die Qualität der Algorithmen getroffen werden, da das Einsatzgebiet der zu untersuchenden Algorithmen die Vorverarbeitung in automatischen Spracherkennungssystemen ist.

Zur Messung der Worterkennungsrate wird, wenn nicht anders angegeben, ein Zahlenerkenner auf Basis des HTK-Softwarepaketes [HTK03] benutzt, der die englischen Zahlen eins bis neun, sowie “oh” und “zero” erkennt. Er wurde mit dem unbearbeiteten Trainingsset der TIdigits-Datenbank trainiert. Die Sprachsignale dieser Datenbank wurden

mit einer Abtastfrequenz von 20.000 Hz aufgenommen.

Als Merkmalsvektoren wurden 12 MFC Koeffizienten, deren erste Ableitung sowie die Energie verwendet. Die Merkmale wurden alle 10 ms anhand von 16 ms langen Fenstern ermittelt. Pro Ziffer wurden 18 Zustände trainiert.

Die Ermittlung der Worterkennungsrate wurde immer mit allen Testdaten der TIdigits-Datenbank durchgeführt. Die Worterkennungsrate mit den unbearbeiteten Testdaten betrugt 98,94%.

Es ist anzumerken, dass es sich hierbei um einen Worterkenner mit einer äußerst hohen Worterkennungsrate handelt, da – z.B. im Gegensatz zu Spracherkennungssystemen für kontinuierliche Sprache – nur sehr wenige Worte unterschieden werden müssen und zudem eine sehr hohe Zustandszahl trainiert wurde. Deswegen reagiert er auch nicht so anfällig auf Störungen (Störgeräusche, Hall) wie jene.

Wegen der fehlenden Grammatik können aber keine fehlerkorrigierenden Maßnahmen im Anschluß an die Worterkennung durchgeführt werden, so dass geringe Verbesserungen des Signals, die normalerweise bei Spracherkennungssystemen für kontinuierliche Sprache zu deutlichen Verbesserungen der WER führen, hier nur geringe Verbesserungen der WER bringen.

5 Automatische Spracherkennungssysteme und Hall

Zunächst werden Untersuchungsergebnisse vorgestellt, die zeigen, wie sich Hall auf die Worterkennungsrate von automatischen Spracherkennungssystemen auswirkt.

Zum Bestimmen des Einflusses von Hall auf die Ergebnisse von Spracherkennungssystemen wurden die reinen Testdaten mit gemessenen Raumimpulsantworten verschiedener Räume gefaltet.

Raum	Größe des Raumes [m]	Abstand Quelle-Mikrofon [m]	Nachhallzeit [sec]	Systemabstand	WER [%]
reines Signal				1,0	98,9
Hallarmer Raum	2 x 2 x 2	0,5	0,03	7,3	98,18
Büroraum	3,5x2x3,5	1,0	0,19	51,6	79,31

Tabelle 1: Vergleich Auswirkung von Hall auf WER

Die WER wurde mit dem bereits beschriebenen System ermittelt. Die Testergebnisse sind in Tabelle 1 und Abbildung 2 zusammengefasst. Als Maßzahlen für den Hall wurden der Systemabstand und die Nachhallzeit angegeben.

Die Grafik zeigt wie erwartet, dass starker Hall (Büroraum) die Worterkennungsrate stärker negativ beeinflusst als schwacher Hall (hallarmer Raum). So sinkt die Worterkennungsrate im Büroraum um ca. 20 Prozentpunkte dagegen im hallarmen Raum nur um knapp einen Prozentpunkt.

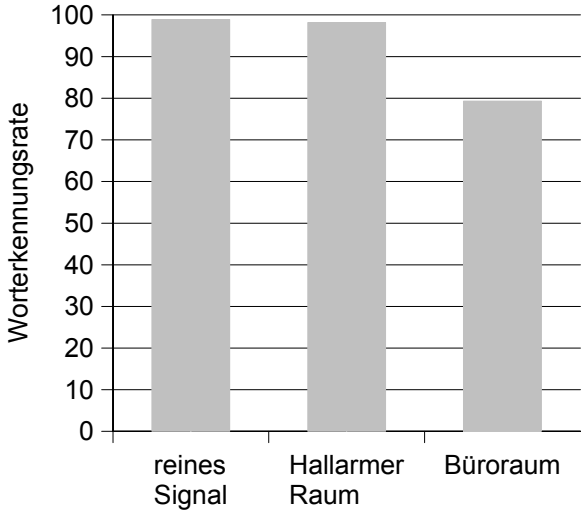

Abbildung 2: Vergleich Auswirkung Hall auf WER

In einem weiteren Versuch war zu ermitteln, welcher Bereich des Halls sich besonders negativ auf die Worterkennungsrate auswirkt. Dazu wurden die gemessenen

Raumimpulsantworten an verschiedenen Stellen abgeschnitten, mit den Testdaten gefaltet und anschließend die Worterkennungsraten gemessen. Die Messergebnisse sind in Abbildung 3 dargestellt.

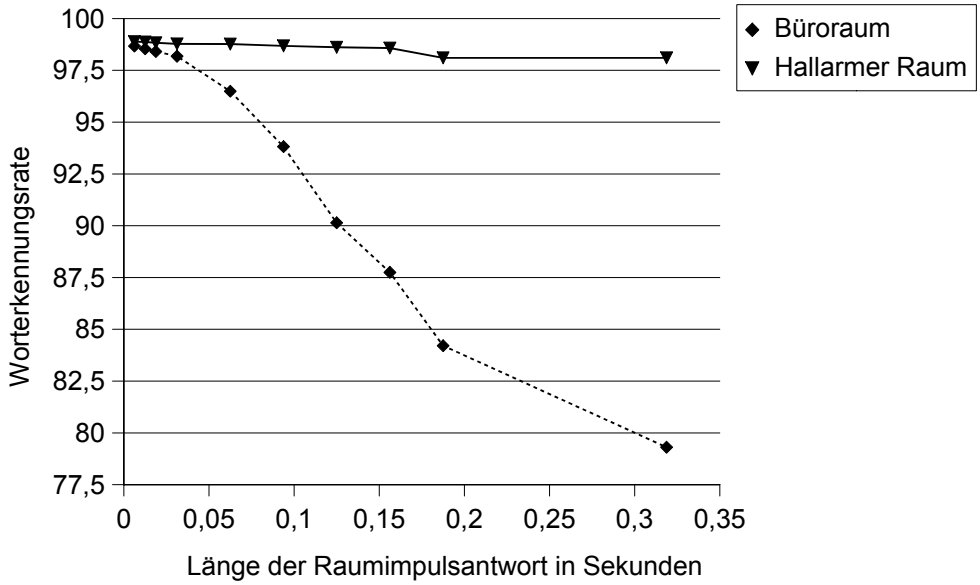

Abbildung 3: Einfluss des Halls auf WER

Im Einklang mit den Ergebnissen von [Sel03] hat sich besonders der lange Nachhall als störend erwiesen. Während sich die Verluste durch die ersten Echos bis ca. 50 ms als moderat erwiesen haben, verursachen die späteren Echos enorme Verluste in der Worterkennungsrate. Die moderaten Verluste am Anfang lassen sich dadurch erklären, dass der Spracherkenner eine Fensterlänge von 16 ms besitzt. Die erste Ableitung der MFC-Koeffizienten, die ebenfalls zum Merkmalsvektor gehören, werden über Differenzbildung der MFC-Koeffizienten zweier benachbarter Fenster gebildet, so dass der Merkmalsvektor anhand von 26 ms Sprachsignal ermittelt wird. Innerhalb dieses Zeitbereichs ist es dem Spracherkenner möglich, Halleinflüsse zu unterdrücken.

Als Ergebnis dieser Testreihen kann man zusammenfassen, dass es besonders den langen Nachhall zu bekämpfen gilt.

6 Algorithmen der Enthallung

Die betrachteten Lösungsvorschläge lassen sich in 3 Gruppen zusammenfassen:

- Delay-and-Sum-Beamformer
- Cepstrum-basierter Ansatz
- Adaptive Verfahren

6.1 Der Delay-and-Sum-Beamformer

Beim Prinzip des Delay-and-Sum-Beamformers wird das Sprachsignal mit mehreren Mikrofonen aufgezeichnet und kohärent aufaddiert.

Bei diesem Lösungsansatz wird ausgenutzt, dass die Raumimpulsantworten zweier Mikrofone ab einer gewissen Verzögerung unkorreliert sind. Außerdem wird angenommen, dass ihr Mittelwert dann Null ist. Verzögert man die Mikrofonsignale, so dass sich die direkten Pfade, also die Nutzsignalanteile, kohärent, die Echos jedoch inkohärent überlagern, erreicht man, dass sich das Nutzsignal-zu-Echoverhältnis um etwa 3 dB verbessert.

Der Sum-and-Delay-Beamformer ist ein bereits gut erforschter Lösungsansatz. Auf seiner Basis stehen zur Ermittlung der benötigten Verzögerungen und Verstärkungen effiziente Algorithmen zur Verfügung.

In den hier durchgeführten Untersuchungen wurden Raumimpulsantworten verwendet, die von einer bekannten Quelle zu verschiedenen Mikrofonen gemessen worden sind. Die kohärente Überlagerung wurde so simuliert, dass die Maxima der Absolutwerte der einzelnen Raumimpulsantworten ermittelt, die Raumimpulsantworten entsprechend verzögert und aufaddiert wurden. Es wurde darauf geachtet, dass die Maxima jeweils positives Vorzeichen besaßen, andernfalls wurde die betreffende Impulsantwort mit dem Wert -1 multipliziert. Mit der so ermittelten Impulsantwort wurden anschließend die Testdaten gefaltet und die Spracherkennungsperformance ermittelt. Dieses Vorgehen bei der Simulation ist berechtigt, da sowohl Faltung als auch die Addition lineare Operationen sind, deren Reihenfolge vertauscht werden darf.

In Abbildung 4 wurde der Systemabstand der resultierenden Impulsantworten aufgetragen und mit den einkanaligen Impulsantworten verglichen. Sie zeigt, dass keine Verbesserung erreicht wurde. Auch konnte in den Hörproben keine Verbesserung erkannt werden. Es

war also nicht zu erwarten, dass die WER gesteigert werden kann.

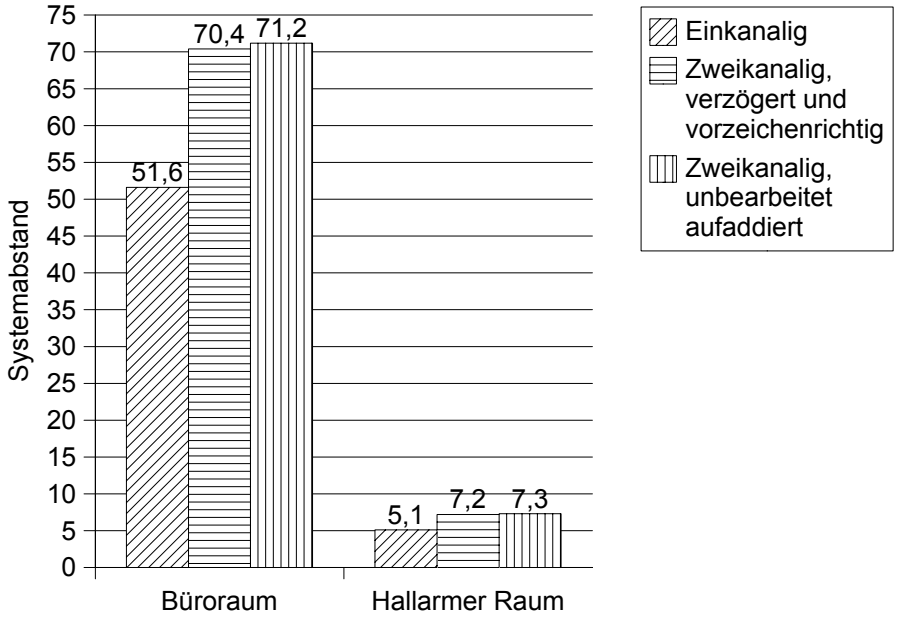

Abbildung 4: Systemabstand der (Raum-)Impulsantworten

Wie man der Abbildung 5 entnehmen kann, wird diese Erwartung durch die experimentelle Bestimmung der Worterkennungsraten bestätigt. So wird in den hier betrachteten Szenarien kein Gewinn gegenüber der Verwendung eines Mikrofons (einkanalig) erzielt. Im Gegenteil: In einigen Fällen musste ein Verlust von bis zu 2 Prozentpunkten beobachtet werden.

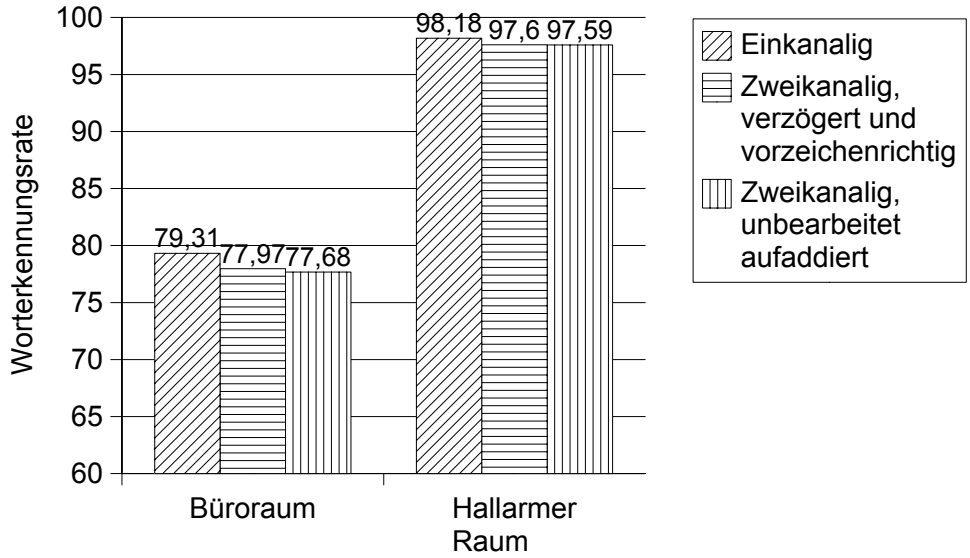

Abbildung 5: Vergleich Verbesserung der WER wegen Hallreduzierung mittels Beamformer

Es wird deshalb vorgeschlagen, dass vor Anwendung dieses Verfahrens überprüft wird, ob

es nicht möglich ist, das beste Mikrofonsignal auszuwählen und dieses für die Spracherkennung zu verwenden.

Es sollte beachtet werden, dass hier im Vergleich zu allen möglichen Raumimpulsantworten nur eine Stichprobe von sehr geringem Umfang untersucht wurde, so dass die Aussage nicht unbedingt repräsentativ ist und in anderen Untersuchungen [SEL03] durchaus, wenn auch nur geringe, Verbesserungen in der Worterkennungsrate erzielt worden sind.

6.2 Cepstrum-basierter Ansatz

6.2.1 Mathematische Herleitung

Wie bereits oben gezeigt, kann Hall als Faltung eines Sprachsignals s(t) mit einer Raumimpulsantwort h(t) beschrieben werden.

$$x(t) = s(t) * h(t) \tag{4}$$

Aus der Systemtheorie ist bekannt, dass mit Hilfe der Fouriertransformation die Faltung in ein Produkt umgewandelt werden kann.

$$X(j\omega) = S(j\omega) \cdot H(j\omega) \tag{5}$$

Eine anschließende Logarithmierung der Gleichung überführt die Multiplikation in eine Addition

$$\ln(X(j\omega)) = \ln(S(j\omega) \cdot H(j\omega)) = \ln(S(j\omega)) + \ln(H(j\omega)) \quad , \tag{6}$$

wobei eine nachfolgende inverse Fouriertransformation nichts an dieser Tatsache ändert

$$F^{-1}\{\ln(X(j\omega))\} = F^{-1}\{\ln(S(j\omega))\} + F^{-1}\{\ln(H(j\omega))\} \quad . \tag{7}$$

Die so erhaltenen Terme sind in der Literatur als Cepstren bekannt. Verkürzt lässt sich die Gleichung 7 somit schreiben als

$$X = S + H \quad . \tag{8}$$

Wie gezeigt wurde, kann somit die Faltung im Zeitbereich in eine Addition im Cepstralbereich übergeführt werden. Bei sich nicht verändernder Umgebung ist Hall ein Offset, der subtraktiv entfernt werden kann.

Der Offset kann durch eine Mittelwertbildung über das Cepstrum verschobener Zeitfenster geschätzt werden.

$$H_{mean} = \frac{1}{N} \sum_{n=1}^{N} X_n \tag{9}$$

Natürlich wird dabei auch der cepstrale Mittelwert der Sprache dem Hall zugeordnet und später entfernt. Da jedoch ein Großteil der heute eingesetzten ASR-Systeme ebenfalls während der Vorverarbeitung den cepstralen Mittelwert vom Signal abziehen, ist durch die

Subtraktion des cepstralen Mittelwerts keine Einbuße bei der WER zu erwarten.

Oben wurde festgestellt, dass bereits bei ASR-Systemen der cepstrale Mittelwert entfernt wird. Somit stellt sich die Frage: Wozu sollte man diesen Vorgang in einem vorgelagerten Schritt nochmals durchführen und warum sollte dies eine Verbesserung der WER nach sich ziehen? Der Grund liegt darin, dass bei ASR-Systemen die Fensterlänge, auf der diese Operation angewandt wird, nur sehr kurz ist. Bei dem hier trainierten System beträgt sie nur 16 ms. D.h., es kann auf diese Weise nur der Nachhall dieser Länge entfernt werden. Raumimpulsantworten in Büroräumen sind jedoch viel länger. Diesem kann man nur gerecht werden, wenn die Fensterlänge diesem Wert anpasst wird, so dass eine große Fensterlänge für die Hallunterdrückung erforderlich ist und eine weitere, aber kürzere für die Spracherkennung.

Für das Entfernen von Hall wurden Fensterlängen zwischen 0,5 und 2 Sekunden gewählt. Die Wahl wird zum einen dadurch begründet, dass Hall in Büroräumen selbst äußerst lang ist (ca 200 ms). Das Fenster sollte mindestens diesen Bereich abdecken, um den Einfluss von Hall zu erfassen. Zum anderen gibt es Ein- und Ausschwingartefakte, die durch die große Fensterlänge kompensiert werden sollen. So befindet sich z.B. im Signal am Anfang eines Fensters noch der Hall des vorhergehenden Fensters. Ferner ist der Algorithmus störunempfindlicher gegenüber Sprachpausen, in denen die Bildung des Logarithmus wegen fehlenden Signals nicht möglich ist. Bei großen Fensterlängen müssen somit kurze Sprachpausen, die z.B. zwischen einzelnen Worten vorkommen, nicht detektiert werden. Die Implementierung wird vereinfacht.

Ein weiteres Problem, das bei der Rekonstruktion auftritt, sind zusätzliche Signalspitzen, die an den Fenstergrenzen entstehen. Diese Artefakte lassen sich dadurch reduzieren, dass bei der Fensterung Fensterfunktionen, wie Gauss- oder Kaiserfenster, benutzt werden. Des weiteren sollten sich die benachbarten Fenster überlappen, so dass die Artefakte, die an den Fenstergrenzen entstehen, herausgemittelt werden können. Bei einem Überlappungsverhältnis von 127:128 wurden die Peaks im rekonstruierten Signal in Hörversuchen nicht mehr festgestellt.

Die Kombination mehrerer Mikrofonsignale kann durch kohärente Überlagerung der rekonstruierten Signale erfolgen.

6.2.2 Implementierung

Der Algorithmus wurde als C++ Programm implementiert. Ferner wurde, um Zeit und Arbeitsspeicher zu sparen, die online Variante gewählt, da es bei diesem Ansatz aufgrund der besonders starken Sprecherabhängigkeit des Cepstrums vollkommen unmöglich ist, das für einen Sprecher ermittelte mittlere Cepstrum auf das Sprachsignal eines anderen Sprechers anzuwenden.

Das Struktogramm des Programms ist in Abbildung 24 dargestellt.

```
c_mean[k] = 0
w[k]=Gaussfenster(Fensterlänge)
Werte x[k] des nächsten Fensters ermitteln
For k=1 to Fensterlänge
    x[k]=w[k]*x[k]
X=fft(x)
Unwrap(X)
C=ifft(log(abs(X))+iarg(X))
For Q=1 to Fensterlänge
    c_mean[Q]=(1-alpha)*c_mean[Q]+
                        alpha*C[Q]
    C[Q]=C[Q]-c_mean[Q]
Y=exp(fft(C))
Y=ifft(Y)
y=AGC(y)
Solange bis Ende des Signals erreicht wird
```

Abbildung 6: Struktogramm des Cepstrum-basierten Algorithmus

Das Struktogramm besitzt einige Besonderheiten, die in der mathematischen Herleitung übergangen worden sind.

Wendet man auf eine komplexe Zahl den Logarithmus an, enthält der Realteil des zurückgegebenen Wertes den reellen Logarithmus des Betrages, der Imaginärteil hingegen die Phase der komplexen Zahl.

$$\log(z) = \log(r \cdot e^{i\phi}) = \log(r) + i\phi \qquad (10)$$

Der Realteil ist somit eindeutig definiert. Die Phase hingegen ist mehrdeutig. Eine Addition von Vielfachen von 2 π ändert nichts an der zugrunde liegenden komplexen Zahl. Um dieses Problem zu umgehen, wird in den meisten mathematischen Bibliotheken die Phase auf das Interval]-π; π] abgebildet.

Für die Fouriertransformierte bedeutet dies, dass ihre so berechnete Phase Sprünge aufweist, wenn die Phase größer als π oder kleiner als -π wird. Bei der späteren Berechnung der inversen Fouriertransformierten zur Bestimmung des Cepstrums würden somit viele Artefakte entstehen, die nicht im Signal vorhanden waren. Aus diesem Grund wurde der Schritt des Unwrappings nach der Logarithmusberechnung eingefügt. In diesem Schritt wird die Differenz der Phasen zweier aufeinander folgender Frequenz-Bins ermittelt. Ist die Differenz größer als π, werden auf alle nachfolgenden Werte 2 π addiert oder subtrahiert.

Eine weitere Besonderheit stellt die Ermittlung des Mittelwertes dar. Sie erfolgt rekursiv gemäß der Formel [Nie03]:

$$C(m) = \alpha C(m-1) + (1-\alpha) X(m) \quad . \tag{11}$$

Der Gewichtsfaktor α wurde zu 0.8 gewählt.

Vor der Ausgabe des rekonstruierten Signals wurde zusätzlich eine einfache automatische Lautstärkeregelung implementiert, die deswegen nötig ist, weil Signalelemente am Anfang einer Sprachdatei nur durch Überlagerung weniger Fenster gemittelt werden, in der Mitte des Sprachsignals jedoch durch die Überlagerung vieler (z.B. 128), so dass sich starke Pegelschwankungen ergaben.

6.2.3 Versuchsergebnisse

Die Ermittlung der WER erfolgte dadurch, dass das beschriebene Programm auf alle Dateien des Testdatensatzes angewandt wurde, nachdem diese mit den gemessenen Raumimpulsantworten gefaltet worden waren. Anschließend wurde die WER mit dem bereits beschriebenen ASR-System bestimmt.

Abbildung 7 zeigt, dass bei diesem Verfahren leichte Gewinne bei der WER erzielt werden konnten. Allerdings standen zur Adaption keine 10 Sekunden sondern jeweils nur zwischen 2 und 6 Sekunden Sprache, d.h. nur wenige cepstrale Mittelwerte, zur Verfügung und es kann nicht davon ausgegangen werden, dass die Adaption schon ihr Maximum an Enthallungsfähigkeit erreicht hatte.

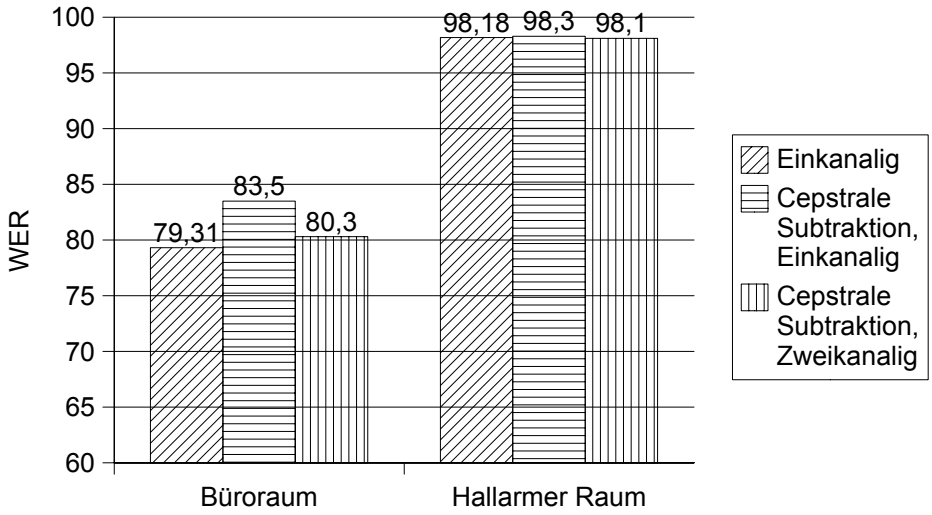

Abbildung 7: Vergleich Verbesserung der WER aufgrund der Hallreduzierung mittels cepstraler Subtraktion

Ferner dürfte bei längeren Testsätzen die WER weiter steigen, weil der Zeitbereich, in dem die Ausgangsdaten über nur wenige Fenster gemittelt werden, prozentual abnimmt.

6.3 Adaptive Verfahren unter Verwendung von Statistik zweiter Ordnung

6.3.1 Grundprinzip

Die dritte Gruppe der in dieser Arbeit untersuchten adaptiven Verfahren zur Verringerung von Hall beruhen auf Auswertungen von Statistiken zweiter Ordnung. Sie sind vor allem im Hinblick auf zeitlich veränderliche akustische Umgebungen interessant [Hay96].

Das Prinzip effizienter Sprachkodierer basiert auf der Annahme, dass Sprache als autoregressiver Rauschprozeß modelliert werden kann, bei dem das Filter des Prozesses je nach Abtastrate und Modell eine Länge von 10 bis 100 Filterelementen besitzt und sich zeitlich relativ langsam (alle 20-50 ms) ändert.

Da das anregende Signal als weißes Rauschen angenommen wird, dessen Autokorrelationsfunktion einen Peak bei einer Verzögerung von Null aufweist, ansonsten jedoch Null ist, ergibt sich unter Beachtung der Faltungsgesetze für Korrelation [Gir97], dass die Autokorrelationsfunktion von Sprachsignalen nur bis Verzögerungen um die 100 Abtastwerte signifikant von Null verschiedene Werte aufweist.

Nun wurde bei der Behandlung von Hall festgestellt (s. oben), dass er, mathematisch gesehen, durch eine Faltung des Sprachsignals mit einer Raumimpulsantwort entsteht.

Ferner wurde gezeigt, dass Raumimpulsantworten sehr lang sind, so dass zusätzlich zu den bereits vorhanden Korrelationen im Sprachsignal weitere hinzu kommen, die auch bei großen Verzögerungen bestehen. Die im Folgenden beschriebenen Algorithmen versuchen Filter zu finden, mit deren Hilfe diese Korrelationen verringert werden können, was gleichzeitig einer – zumindest teilweisen – blinden Enthallung gleichkommt.

6.3.2 Der Algorithmus von Gillespie und Atlas

6.3.2.1 Mathematische Herleitung

Der erste Algorithmus dieser Gruppe wurde von Gillespie und Atlas entwickelt [Gil03]. Abbildung 8 zeigt die Systemkonfiguration.

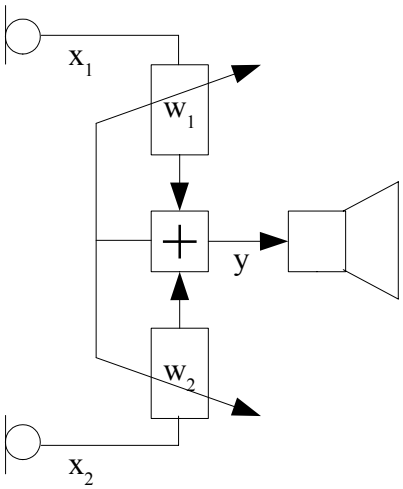

Abbildung 8: Systemkonfiguration des Algorithmus von Gillespie und Atlas

Am Mikrofon l liegt das Signal $x_l[k]$ an, das sich aus der Faltung des Quellsignal s[k] mit der Raumimpulsantwort $h_l[k]$ ergibt:

$$x_l[k] = s[k] * h_l[k] \quad . \tag{12}$$

Für das Ausgangssignal des Systems y[k] gilt,

$$y[k] = \sum_l x_l[k] * w_l[k] \quad , \tag{13}$$

wobei $w_l[k]$ die Impulsantwort des adaptiven Filters für das Mikrofonsignal $x_l[k]$ ist. Die Autokorrelierte des Ausgangssignals lautet

$$R_{yy}[\kappa] = \sum_k y[k] \cdot y[\kappa - k] \quad . \tag{14}$$

Als zu minimierende Kostenfunktion wird

$$\epsilon[\kappa] = \beta[\kappa] \cdot (R_{yy}[\kappa] - R_{dd}[\kappa])^2 \tag{15}$$

angegeben, mit R_{dd} als angestrebte Autokorrelierte des Ausgangssignal und β als eine Gewichtsfunktion, die den Einfluß der Komponenten der Autokorrelierten auf die Kostenfunktion wichtet.

Die Minimierung erfolgt durch die Wahl der Filterkoeffizienten $w_l[k]$, so dass ein Optimum dann erreicht ist, wenn

$$\frac{\partial \epsilon[\kappa]}{\partial w_l[k]} = 0 \tag{16}$$

gilt.

Führt man die Ableitung aus, erhält man

$$\frac{\partial \epsilon[\kappa]}{\partial w_l[k]} = \frac{\partial \beta[\kappa](R_{yy}[\kappa] - R_{dd}[\kappa])^2}{\partial w_l[k]} \quad . \tag{17}$$

Da nur R_{yy} von den Filterkoeffizienten abhängt, kann die Gleichung 17 in

$$\frac{\partial \epsilon[\kappa]}{\partial w_l[k]} = 2 \cdot \beta[\kappa](R_{yy}[\kappa] - R_{dd}[\kappa]) \frac{\partial R_{yy}[\kappa]}{\partial w_l[k]} \tag{18}$$

umgeformt werden. Mit der Beziehung

$$\frac{\partial R_{yy}[\kappa]}{\partial w_l[k]} = R_{x_l y}[k - \kappa] + R_{x_l y}[k + \kappa] \tag{19}$$

lässt sich die Formel 18 in

$$\frac{\partial \epsilon[\kappa]}{\partial w_l[k]} = 2 \cdot \beta[\kappa](R_{yy}[\kappa] - R_{dd}[\kappa])(R_{x_l y}[k - \kappa] + R_{x_l y}[k + \kappa]) \tag{20}$$

umwandeln.

Für den Gradienten dw_l des Rekonstruktionsfilters des l-ten Mikrofonsignals folgt damit

$$\nabla w_l[k] = \sum_{\kappa} \beta[\kappa](R_{yy}[\kappa] - R_{dd}[\kappa])(R_{x_l y}[k - \kappa] + R_{x_l y}[k + \kappa]) \quad . \tag{21}$$

Dieser Ausdruck vereinfacht sich zu

$$\nabla w_l[k] = \sum_{\kappa > c} (R_{yy}[\kappa])(R_{x_l y}[k-\kappa] + R_{x_l y}[k+\kappa]) \quad , \tag{22}$$

wenn die gewünschte Autokorrelierte R_{dd} ab einer Verzögerung von c Abtastwerten (Schutzbereich) 0 sein soll, und die Gewichtsfunktion β ab diesem Wert 1 beträgt.

Gillespie und Atlas schlagen zudem eine Normierung nach dem Schema des NLMS [Hay96] Algorithmus vor, so dass die Updategleichung folgendes Aussehen erhält:

$$w_l[k] = w_l[k] - \mu \frac{\nabla w_l[k]}{\sqrt{\sum_k \sum_l \nabla w_l[k]^2}} \quad . \tag{23}$$

6.3.2.2 Implementierung

Dieser Algorithmus wurde in Matlab als Offline-Algorithmus implementiert. Aus den Formeln kann direkt das in Abbildung 9 angegebene Struktogramm abgeleitet werden. Für die Schätzung der Autokorrelierten wurde sowohl die „biased“ als auch die „unbiased“ Variante verwendet. Allerdings ließen sich keine nennenswerte Unterschiede in den Ergebnissen feststellen. Die nachfolgenden Testversuche wurde deshalb nur mit der „unbiased“ Version durchgeführt.

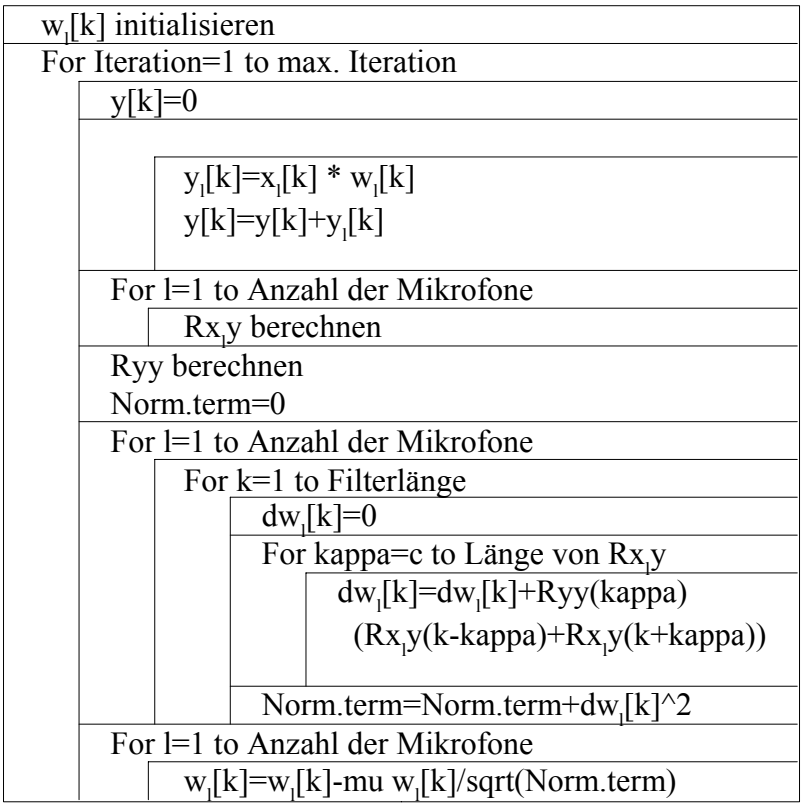

Abbildung 9: Struktogramm des Algorithmus von Gillespie und Atlas

6.3.2.3 Versuchsergebnisse

In den Versuchen wurden reine Sprachsignale verschiedener Länge mit gemessenen Raumimpulsantworten gefaltet, um so die Mikrofoneingänge eines Mikrofonarrays zu simulieren. Die Raumimpulsantworten wurden vorher jedoch schon derart verzögert, dass die Direktpfade zeitgleich die verschiedenen Mikrofone erreichen. Dies kann bei einer späteren Implementierung dadurch erreicht werden, dass dem Algorithmus der Teil des Delay-and-Sum-Beamformers vorgeschaltet wird, der die Laufzeiten der Direktpfade ausgleicht.

Zur Ermittlung der WER wurde nach abgeschlossenem Training die Gesamtimpulsantwort i[k] von Sprecher bis Systemausgang durch Faltung der einzelnen Raumimpulsantworten mit den Systemimpulsantworten berechnet

$$i[k] = \sum_l h_l[k] * w_l[k] \quad . \tag{24}$$

Anschließend wurde der komplette Testdatensatz mit der so ermittelten Gesamtimpulsantwort gefaltet und die WER mit dem bereits beschriebenen ASR-System

ermittelt.

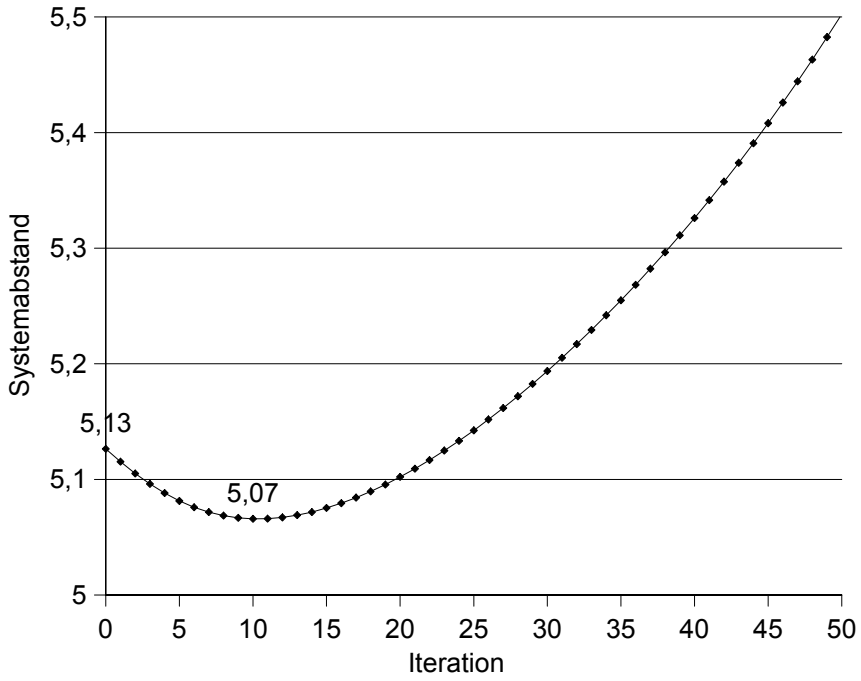

Abbildung 10: Systemabstand der Gesamtimpulsantwort über Iteration im hallarmen Raum, Länge des Rekonstruktionsfilters 1024, Länge des Schutzbereiches 32, Abtastrate 20.000 Hz

Das Ergebnis zeigt, dass der Systemabstand in Abhängigkeit der Iteration (Abbildung 10) bei niedriger Iterationszahl abnimmt, ein Minimum erreicht und anschließend jedoch wieder kontinuierlich ansteigt.

Die WER wurden mit dem Filter ermittelt, mit dem das Minimum des Systemabstandes erreicht wurde. Ein Vergleich (Tabelle 2) mit der des ungefilterten Mikrofonsignals zeigt, dass kein Gewinn sondern ein Verlust erzielt wurde. Erfolglos blieb auch der Versuch, das Rekonstruktionsfilter zu bestimmen, wenn vor dem Adaptionsalgorithmus eine lineare Prädiktion durchführte, um so die Korrelation zu entfernen, die durch den Sprachtrakt des Menschen dem Signal hinzugefügt worden ist.

Länge der Rekonstruktionsfilter	Länge des Schutzbereiches	Büroraum		Hallarmer Raum	
		WER	Systemabstand	WER	Systemabstand
Vergleichswert, einkanalig, unbearbeitet		79,3	51,6	98,2	5,1
256	8	76,2	48,5	97,6	5,0
256	24	76,5	47,9	97,6	5,0
256	32	76,6	47,7	97,5	5,0
512	8	75,4	48,2	97,6	5,0
512	24	75,3	47,9	97,6	5,0
512	32	75,5	47,8	97,5	5,0
1024	8	74,8	47,8	97,6	5,0
1024	24	74,9	47,6	97,6	5,0
1024	32	75,2	47,5	97,5	5,0

Tabelle 2: WER und Systemabstand der Gesamtimpulsantwort nach Enthallung durch Algorithmus von Gillespie und Atlas, Abtastrate 20.000 Hz, Schrittweite μ = 0,01, LPC

Als problematisch hat sich herausgestellt, dass die Updateterme periodische Komponenten aufweisen (Abbildung 11).

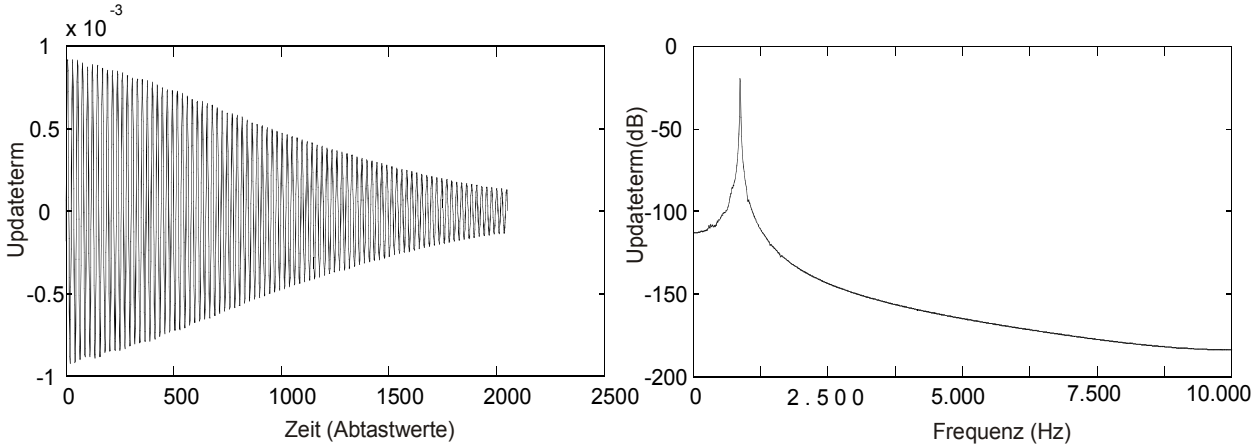

Abbildung 11: Updateterm bei Algorithmus von Gillespie und Atlas nach 10 Iterationen, Länge des Rekonstruktionsfilter 2048, Größe des Schutzbereiches 100, Abtastfrequenz 20.000 Hz, links im Zeitbereich, rechts Fouriertransformierte des Updateterms

Als Resultat dieser periodischen Komponente wird beobachtet, dass das Rekonstruktionsfilter einen stark bandpassartigen Charakter erhält (Abbildung 12), der das Eingangssignal verzerrt und somit eine Spracherkennung erschwert.

Eine genaue Ursache für diese Komponenten konnte nicht ermittelt werden. Durch die persönlichen Hinweise von Herrn Gillespie zur Einstellung der Parameter konnten diese periodischen Komponenten ebenfalls nicht beseitigt werden. Vermutlich handelt es sich um ein Normierungsproblem aufgrund des heuristischen Ansatzes.

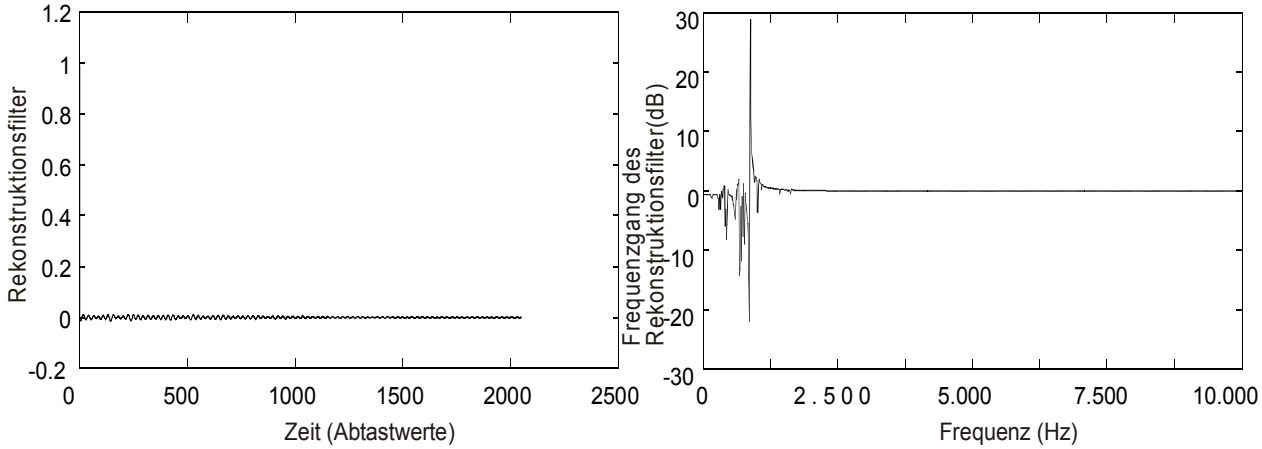

Abbildung 12: Rekonstruktionsfilter bei Algorithmus von Gillespie und Atlas nach 10 Iterationen, Länge des Rekonstruktionsfilter 2048, Größe des Schutzbereiches 100, Abtastfrequenz 20.000 Hz, links im Zeitbereich, rechts Fouriertransformierte des Rekonstruktionfilters

6.3.3 BSS-basierter Ansatz

6.3.3.1 Mathematische Herleitung

Ein weiterer Ansatz, mit dem die Autokorrelationsfunktion ab einer gewissen Verzögerung minimiert werden soll, wurde aus der Blinden Quellentrennung („Blind Source Separation“, BSS) weiterentwickelt. BSS ist ein erfolgreich eingesetztes Verfahren zur Trennung zweier oder mehrerer Signale, die mit mehreren Mikrofonen aufgenommen werden. Die Trennung erfolgt über die Minimierung der Einträge der Kreuzkorrelationsmatrizen der beiden Ausgangssignale des Systems mit Hilfe adaptiver linearer Filterung.

$$\begin{bmatrix} \mathbf{R}_{y1y1} & \mathbf{R}_{y1y2} \\ \mathbf{R}_{y2y1} & \mathbf{R}_{y2y2} \end{bmatrix} \qquad \begin{bmatrix} \mathbf{R}_{y1y1} & 0 \\ 0 & \mathbf{R}_{y2y2} \end{bmatrix}$$

Abbildung 13: Korrelationsmatrix der Systemausgänge bei BSS vor und nach der Trennung (idealisiert)

Die zu minimierende Kostenfunktion für BSS lautet [Buc03]:

$$I(m) = \sum_{i=0}^{m} \beta(i,m)\{\log\det \operatorname{bdiag}\mathbf{R}_{yy}(m) - \log\det \mathbf{R}_{yy}(m)\} \qquad (25)$$

wobei $\mathbf{R}_{yy}(m)$ die Korrelationsmatrix der Ausgänge $\mathbf{y}(m)$ des m-ten Fensters ist:

$$\mathbf{R}_{yy}(m) = \mathbf{Y}^{H}(m)\,\mathbf{Y}(m) \quad . \tag{26}$$

Die Matrizen **Y**(m) beinhalten als Spalten zeitlich sukzessive verschobene Versionen des Block-Ausgangsvektors **y**. Partitioniert man $\mathbf{R}_{yy}$ in Untermatrizen der Größe L x L, wobei L die Fensterlänge zur Bestimmung der Korrelationsmatrix ist, enthalten die Untermatrizen auf der Diagonalen die Autokorrelationsmatrizen und die außerhalb der Diagonalen die Kreuzkorrelationsmatrizen. Der bdiag-Operator bei der BSS lässt genau die Autokorrelationsmatrizen unberührt und setzt die Kreuzkorrelationsmatrizen zu Null.

Die Erweiterung dieses Ansatzes besteht darin, die Autokorrelationsmatrizen der Ausgangssignale in quadratische Untermatrizen zu partitionieren und zusätzlich zu den Kreuzkorrelationsmatrizen jene Untermatrizen zu minimieren, die nicht auf der Diagonalen der Autokorrelationsmatrix liegen. Dazu wird der bdiag-Operator derart abgeändert, dass nur die Untermatrizen der Autokorrelationsmatrizen, die auf der Diagonalen liegen, durch den bdiag-Operator unverändert bleiben und alle anderen Einträge von $\mathbf{R}_{yy}$ durch Null ersetzt werden. Die Größe der nicht veränderten Untermatrizen ist so zu wählen, dass sie der Verzögerung entspricht, bis zu der die natürliche Korrelation der Sprache vorherrscht, die man durch die Filterung nicht beeinträchtigen möchte.

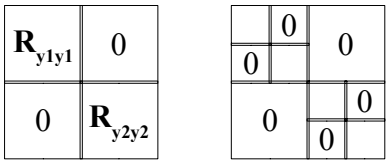

Abbildung 14: Vergleich bdiag Operator bei BSS(links) und Enthallung(rechts): Autokorrelationseinträge werden zusätzlich partitioniert und Nebendiagonalmatrizen auf Null gesetzt

Genauso wie beim Algorithmus von Gillespie und Atlas wird die Kostenfunktion mittels Gradientenabstieg minimiert. Dazu wird die Kostenfunktion (Gleichung 25) nach den Filterkoeffizienten abgeleitet, was nach [Buc03]

$$\frac{\partial \mathrm{I}}{\partial \mathbf{W}^{*}} = 2\sum_{i=0}^{m} \beta(i,m)\,\mathbf{R}_{xy}\left\{\mathrm{bdiag}^{-1}\mathbf{R}_{yy} - \mathbf{R}_{yy}^{-1}\right\} \tag{27}$$

ergibt. Für das Rekonstruktionsfilter **W** wird bei dieser Herleitung die Darstellung als Sylvestermatrix verwendet:

$$\mathbf{W} = \begin{bmatrix} \mathbf{W_1} \\ \vdots \\ \mathbf{W_P} \end{bmatrix} \tag{28}$$

mit

$$\mathbf{W_l} = \begin{bmatrix} w_l[0] & 0 & \cdots & 0 \\ w_l[1] & w_l[0] & \ddots & \vdots \\ \vdots & w_l[1] & \ddots & 0 \\ w_l[L-1] & \vdots & \ddots & w_l[0] \\ 0 & w_l[L-1] & \ddots & w_l[1] \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & w_l[L-1] \\ 0 & \vdots & 0 & 0 \end{bmatrix}, \tag{29}$$

wobei $w_l[k]$ das Rekonstruktionsfilter für das l-te Mikrofonsignal ist.

Die rekursive Adaptionsgleichung für den Gradientenabstieg lautet

$$\mathbf{W}(m) = \mathbf{W}(m-1) - \mu \Delta \mathbf{W}(m) \tag{30}$$

mit $\Delta \mathbf{W}(m)$ als Updateterm und μ als Schrittweite. Mit Hilfe des Konzeptes des natürlichen Gradienten [Buc03] folgt für den Updateterm

$$\Delta \mathbf{W}(m) = \sum_{i=0}^{m} \beta(i,m)\, \mathbf{W}\{\mathbf{R_{yy}} - \mathrm{bdiag}\,\mathbf{R_{yy}}\}\, \mathrm{bdiag}^{-1}\mathbf{R_{yy}} \quad . \tag{31}$$

6.3.3.2 Spezialisierung auf nur ein Ausgangssignal

Um die reine Enthallungsleistung des Algorithmus zu ermitteln, wird im Folgenden eine vereinfachte Konfiguration betrachtet, bei der nur eine Signalquelle vorhanden ist, die mit Hilfe zweier Mikrofone aufgenommen wird. Als Folge dieser Anordnung muss durch das System nur noch ein Ausgangssignal berechnet werden. Dadurch fallen bei der Korrelationsmatrix des Ausgangssignals $\mathbf{R_{yy}}$ die Kreuzkorrelationsterme und der zweite Autokorrelationsterm weg.

Betrachtet man für diesen Fall die Updategleichung genauer, stellt man fest, dass bei gleicher Initialierung der Rekonstruktionsfilter für die verschiedenen Mikrofonsignale identische Filter ermittelt werden: Bei der Multiplikation der Matrix **W** mit dem Ausdruck

	0
0	

Abbildung 15: gewünschte Korrelationsmatrix bei nur einem Ausgangssignal

$\{\mathbf{R}_{yy} - \text{bdiag}\,\mathbf{R}_{yy}\}\,\text{bdiag}^{-1}\mathbf{R}_{yy}$ werden die Untermatrizen $\mathbf{W}_l$ jeweils mit dem eben erwähnten Ausdruck multipliziert. Sind die Matrizen $\mathbf{W}_l$ auf gleiche Weise, z.B. mit dem Einheitsimpuls, initialisiert worden, sind die Updateterme im Fall von nur einem Ausgang für alle Matrizen $\mathbf{W}_l$ identisch und somit auch die neu berechneten Rekonstruktionsfilter w.

Es stellt sich somit heraus, dass dieses Verfahren in diesem Fall in einen Sum-Beamformer und einer dahinter geschalteten einkanaligen Enthallung zerlegt werden kann. Um die Kombination der einzelnen Mikrofonsignale zu verbessern, sollte aus den Überlegungen des Kapitels 6.1 der Sum-Beamformer als Delay-and-Sum Beamformer implementiert werden, damit die möglichen, wenn auch geringen Gewinne, dieses Ansatzes ausgenutzt werden.

Weitere Betrachtungen zum Koeffizientenupdate befinden sich im Kapitel 6.3.5.

6.3.3.3 Implementierung

Die Implementierung der hergeleiteten Gleichungen wurde ebenfalls als Offline-Algorithmus durchgeführt. Zur Beschleunigung der Inversion der Matrix bdiag $\mathbf{R}_{yy}$ wurde die Inversion der Untermatrizen auf der Diagonalen ausgeführt, was nach [Kor68] äquivalent zur Inversion der gesamten Matrix ist. Zudem wurde nur die erste Spalte bzw. die L-te Zeile der Matrix des Rekonstruktionsfilters $\mathbf{W}$ berechnet, da diese alle Einträge des Filters enthalten. Im Falle der ersten Spalte kann die Matrix-Matrix-Multiplikation als Faltung zwischen dem Filter und der ersten Spalte der Matrix, die aus den Autokorrelationsmatrizen berechnet wird, umgeschrieben werden. Im Falle der Berechnung der L-ten Spalte kann die Matrix-Matrix-Multiplikation in eine Vektor-Matrix-Multiplikation umgewandelt werden. Durch diese Maßnahmen ist es möglich, die Rechenzeit um einen Faktor L zu reduzieren.

Als Gewichtsfunktion wurde wegen der Offline-Implementierung die Konstante 1 gewählt. Das Struktogramm der Implementierung ist in Abbildung 16 dargestellt.

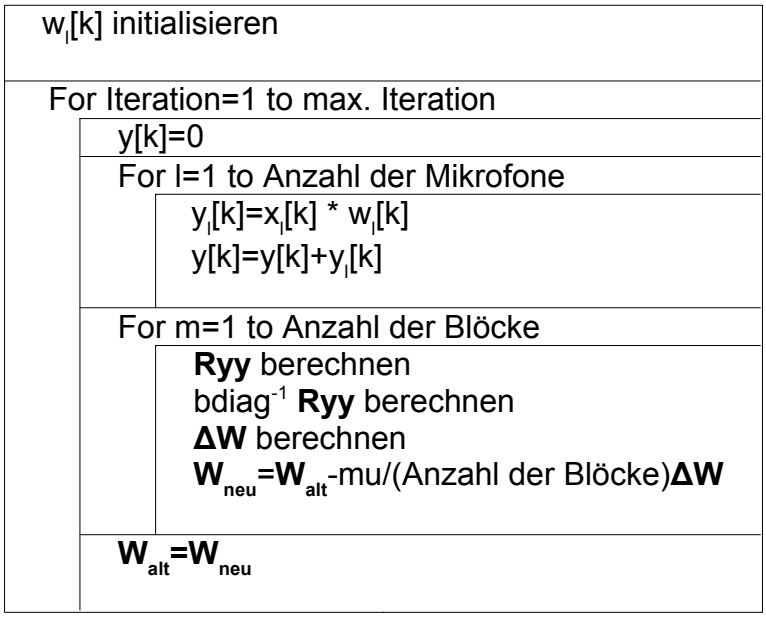

Abbildung 16: Struktogramm des BSS-basierten Algorithmus

6.3.3.4 Versuchsergebnisse

Die Versuchsdurchführung erfolgte analog zur Vorgehensweise beim Ansatz von Gillespie und Atlas. Während der Versuchsdurchführungen wurden sowohl die Filterlänge als auch die Größe der Untermatrizen variiert. Ausserdem stellte sich bei den Untersuchungen

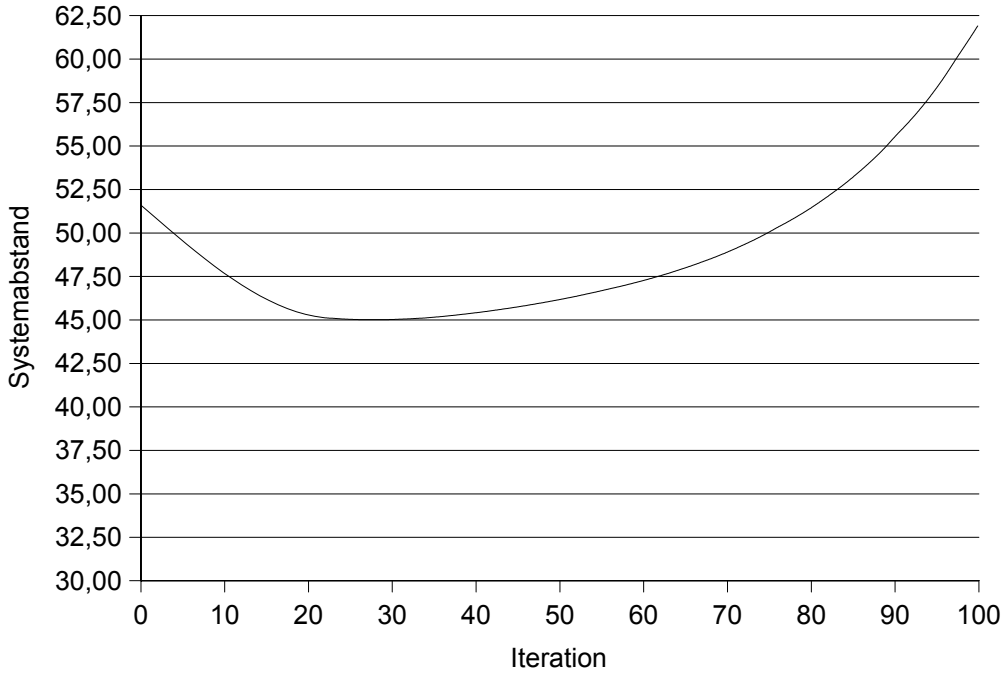

Abbildung 17: Systemabstand der Gesamtimpulsantwort über Iteration im Büroraum, Länge des Rekonstruktionsfilters 512, Größe der Untermatrizen 64x64, Abtastrate 20.000 Hz, Variante 1

heraus, dass die Wahl der ausgewerteten Elemente der Sylvestermatrix **W** einen großen Einfluss auf die Performance hat.

Erste Variante

Zu Beginn der Untersuchung wurden die Updateterme an Hand der ersten Spalte der Filtermatrix **W** bestimmt.

Abbildung 17 zeigt für diesen Fall den Verlauf des Systemabstandes über der Iterationszahl. Die Filterlänge des Rekonstruktionsfilter betrug 512 Elemente, die Größe der Untermatrizen wurde zu 32x32 gewählt. Die Abtastrate des Signals betrug 20.000 Hz. Als Raumimpulsantwort wurde die eines Büroraumes gewählt. Ebenso wie beim Algorithmus von Gillespie und Atlas kann eine Reduzierung des Systemabstandes bei wenigen Iterationen beobachtet werden. Der Systemabstand nimmt aber auch bei diesem Ansatz nach Erreichen eines Minimums wieder zu.

	Größe der Untermatrix		
Filterlänge	**16**	**32**	**64**
64	75,53	68,10	
128	75,81	75,11	64,86
256	76,44	75,95	74,30
512		77,14	74,30
Vergleichswert / Büroraum		79,33	

Tabelle 3: WER bei BSS-basiertem Ansatz (erste Variante) in Abhängigkeit von Länge des Rekonstruktionsfilters und Größe der Untermatrix, Updatewerte wurden an Hand der ersten Spalte der Sylvestermatrix des Filers bestimmt

Tabelle 3 enthält die ermittelten WER. Die Rekonstruktionsfilter wurden so gewählt, dass der Systemabstand bei den jeweils angegebenen Parametern am niedrigsten ist. Während der Versuchsdurchführung wurde die Länge der Rekonstruktionsfilter zwischen den Werten 64 und 512 variiert. Bei längeren Rekonstruktionsfiltern bereitete die Invertierung der Matrizen numerische Probleme, so dass diese Ergebnisse in diesem Vergleich nicht berücksichtigt wurden. Für die Größe der Untermatrizen wurden Werte zwischen 16 und 64 gewählt.

Ähnlich wie beim Algorithmus von Gillespie und Atlas sinkt auch bei dieser Version des BSS-basierten Ansatzes die WER. Der Verlust ist sowohl von der Rekonstruktionsfilterlänge als auch von der Größe der Untermatrizen abhängig. Dabei wird der Verlust mit zunehmender Rekonstruktionsfilterlänge geringer. Bei der Abhängigkeit der WER von der Größe der Untermatrizen spielt offensichtlich das

Verhältnis zwischen der Größe und der Filterlänge eine entscheidende Rolle. Während man bei geringen Längen mit kleineren Untermatrizen die größeren Verluste hat, werden sie mit zunehmender Filterlänge und Untermatrizengröße geringer. D.h., ist das Verhältnis zwischen der Kantenlängen der Untermatrizen und der Filterlänge klein, sind die Verluste größer als bei größeren Verhältnissen.

Doch woher kommt dieser Verlust? Bricht man den Algorithmus ab, wenn das Minimum des Systemabstandes erreicht wurde, und betrachtet dann die Impulsantwort zwischen Signalquelle und dem Ausgang des Enthallungssystem, stellt man fest, dass es dem Algorithmus gelingt, einzelne Peaks, die noch in der Raumimpulsantwort vorhanden waren zu eliminieren. D.h., Echos konnten entfernt werden und die WER müsste eigentlich steigen.

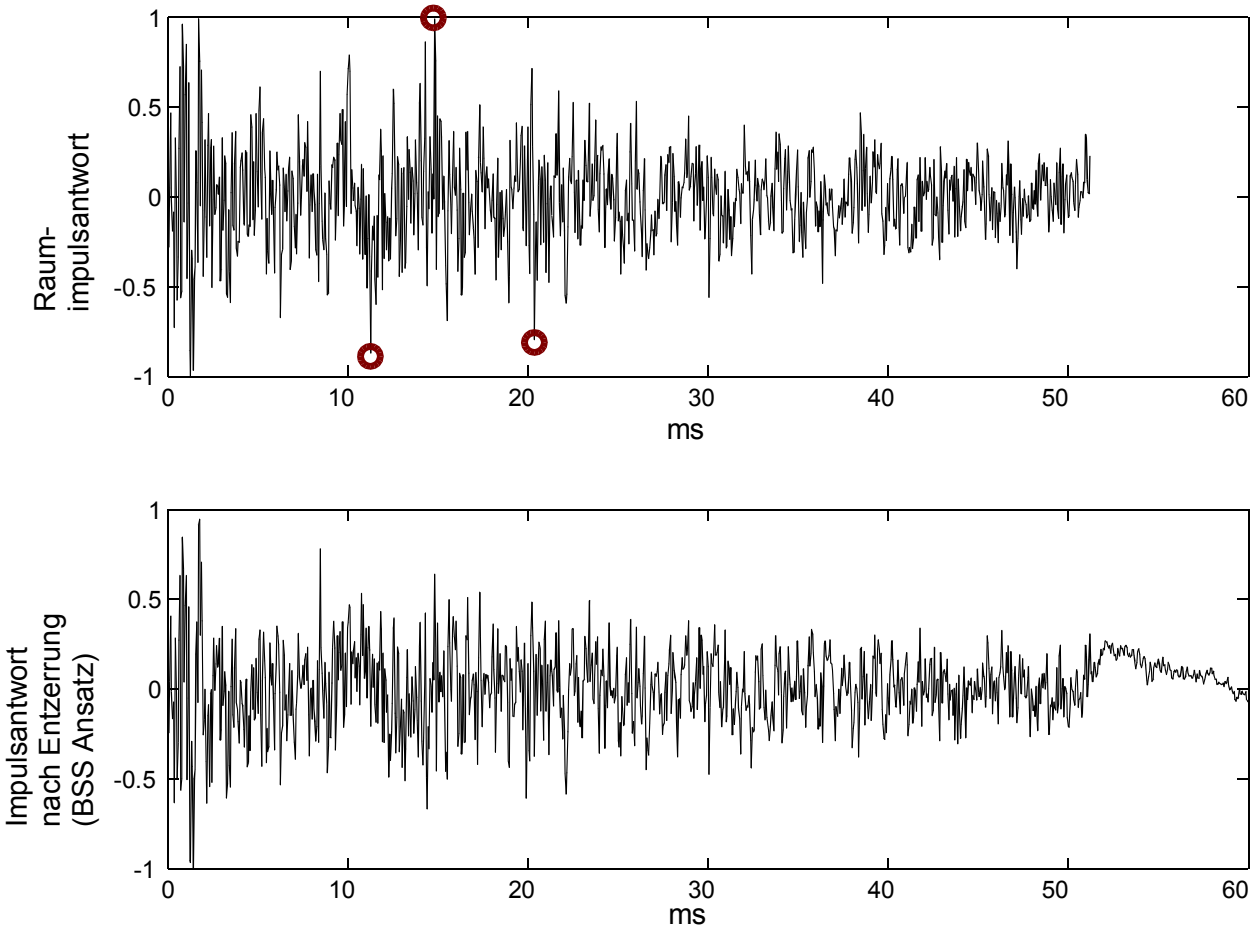

Abbildung 18: Vergleich Raumimpulsantwort vor Enthallung und nach Enthallung, Originalraumimpulsantwort eines Büroraumes, nach 1024 Elementen abgeschnitten,Blocklänge 256, Partitionsgröße 64, 10 Iterationen, eingekreiste Peaks (Echos) konnten beseitigt werden

Ein Hinweis dafür, dass die WER abnehmen könnte, erhält man dagegen aus der Fouriertransformierten der Impulsantwort (Abbildung 19). Während die des Halls in weiten Bereichen einen allpassähnlichen Charakter hat, d.h. das Signal kaum frequenzgefiltert wird, weist das Rekonstruktionsfilter ein stark bandselektives Verhalten auf. Besonders stark ist die Dämpfung im Bereich von 500 bis 1000 Hz. In diesem Bereich ist die Sprachgrundfrequenz angesiedelt.

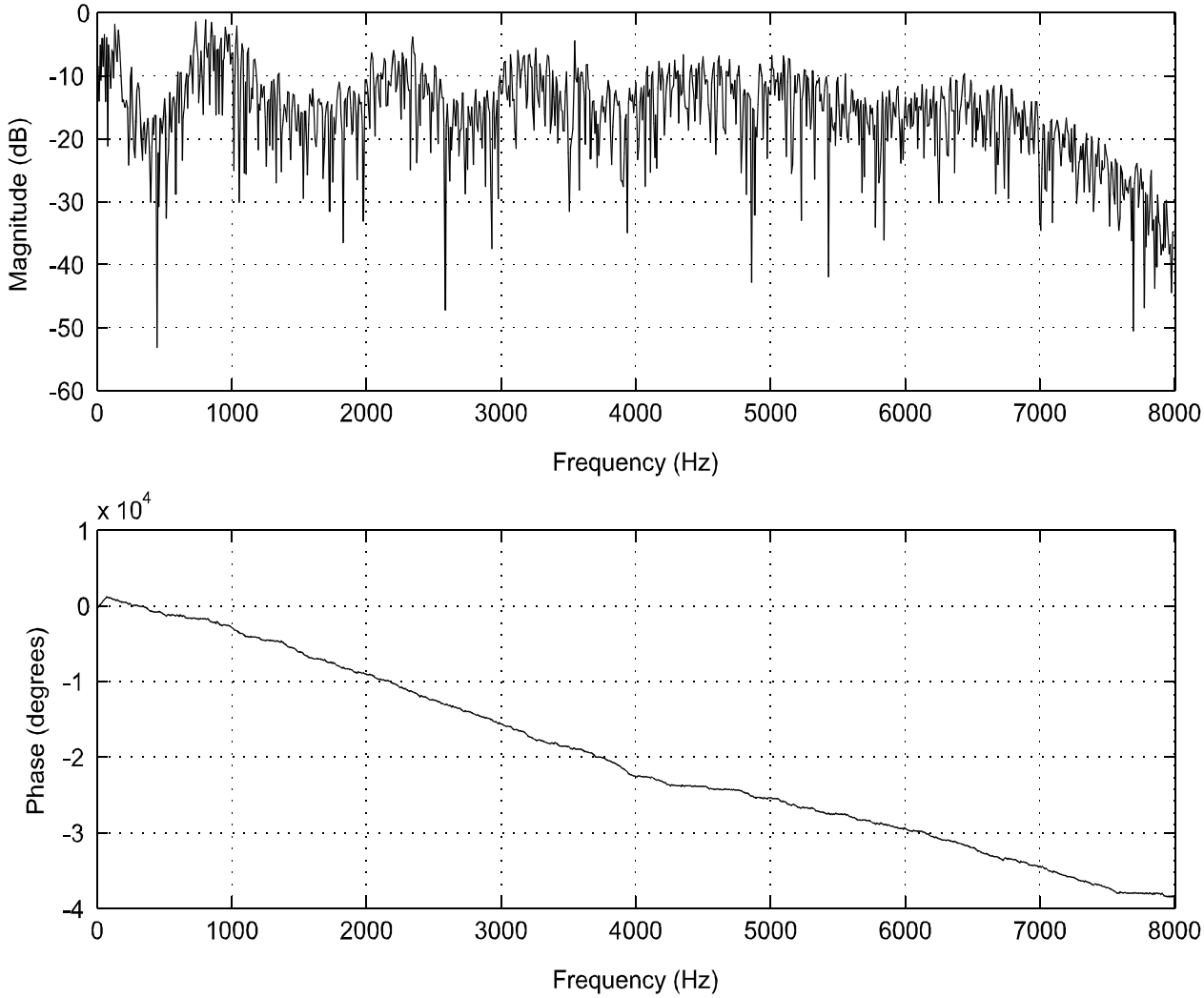

Abbildung 19: Frequenzgang der Raumimpulsantwort aus Abbildung 18 nach Enhallung

In Hörversuchen konnte die daraus resultierende Verzerrung deutlich wahrgenommen werden. Eine Folge dieser Verzerrung war eine reduzierte Verständlichkeit der gesprochenen Worte. Weitergehende Untersuchungen haben gezeigt, dass die Updateterme, ebenso wie beim Algorithmus von Gillespie und Atlas, periodische Komponenten aufweisen, deren Energie mit Iterationszahl zunimmt. Eine Variation der Schrittweite brachte keine Verbesserung. Ein Grund für die periodischen Komponenten

könnte die Normierung in der Updategleichung auf den Term bdiag $\mathbf{R}_{yy}$ sein. Zur Berechnung der neuen Filterkoeffizienten wird nur die erste Spalte der Updategleichung herangezogen. Bei der Normierung wird aber nur die erste Untermatrix der Diagonalen verwendet. Diese Bevorzugung ist jedoch nicht einleuchtend und erklärbar.

Zweite Variante

Aus diesem Grunde wurden für die Filteradaption in einem zweiten Versuch die Werte der L-ten Zeile des Updateterms gewählt. Die Ergebnisse in Tabelle 4 zeigen, dass durch diese Maßnahme sowohl eine Verbesserung des Systemabstandes als auch der WER erreicht werden konnte. Jedoch hat sich auch bei dieser Wahl der Updateterme gezeigt, dass die Updateterme periodische Komponenten aufweisen, so dass die resultierenden Rekonstruktionsfilter zwar keinen Bandpasscharakter dafür aber Hochpasscharakter besitzen (Vgl. Abbildung 20).

Iterationen	**WER**	**Systemabstand**
30	80,24	40,14
50	80,16	40,10
70	80,44	40,09
90	80,50	40,01
Vergleichswert / Büroraum	79,33	51,60

Tabelle 4: WER und Systemabstand bei BSS-basierten Algorithmus (zweite Variante), Länge des Rekonstruktionsfilters 512, Größe der Untermatrizen 32x32

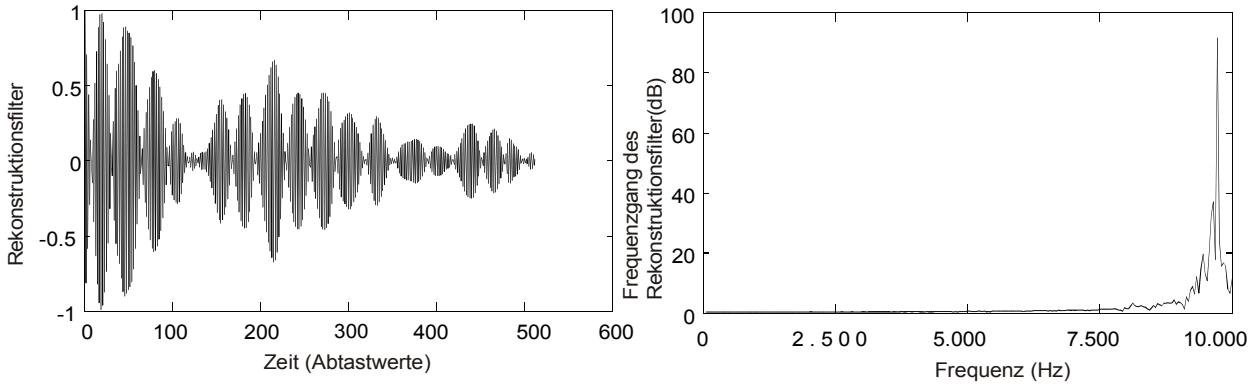

Abbildung 20: Rekonstruktionsfilter bei BSS-basierten Algorithmus (zweite Variante) nach 70 Iterationen, Länge des Rekonstruktionsfilter 512, Größe der Untermatrizen 32x32, Abtastfrequenz 20.000 Hz, links im Zeitbereich, rechts Fouriertransformierte des Rekonstruktionfilters

Eine Begründung für dieses Verhalten liefert die Fouriertransformierte der Raumimpulsantwort (Abbildung 21). Raumimpulsantworten haben einen Tiefpass-ähnlichen Charakter. Durch das Rekonstruktionsfilter muss dieser Frequenzgang

ausgeglichen werden.

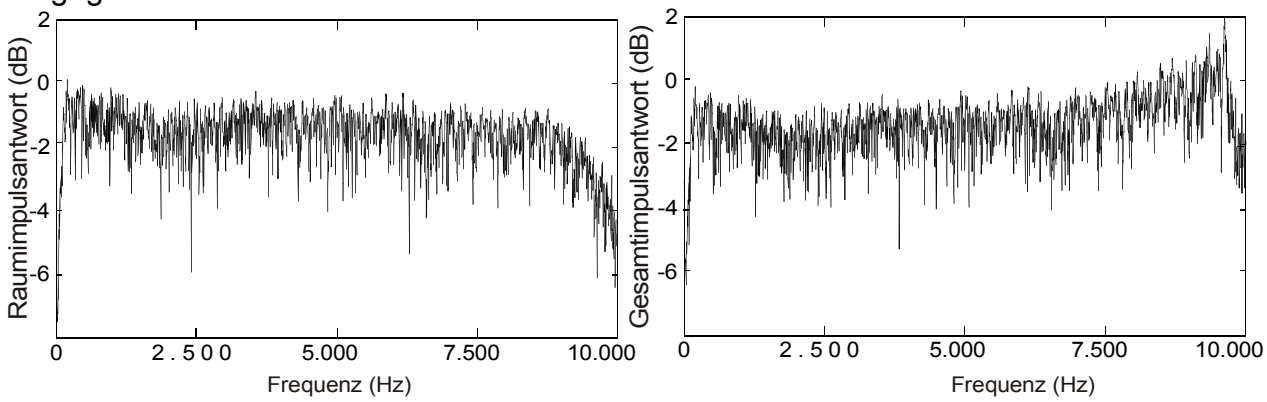

Abbildung 21: links: Frequenzgang der Raumimpulsantwort eines Büroraums, rechts Frequenzgang der Gesamtimpulsantwort (Raumimpulsantwort + Rekonstruktionsfilter)

6.3.4 Prädiktions-basierter Ansatz

6.3.4.1 Mathematische Herleitung

In diesem Ansatz wird das blinde Adaptionsproblem in ein überwachtes Adaptionsproblem überführt. Einer der bekanntesten Algorithmen zur überwachten adaptiven Filterung ist der sog. „Least Mean Squares“ Algorithmus (LMS-Algorithmus) [Hay96]. Der LMS Algorithmus ist ein iteratives Verfahren, adaptiv ein Filter zu bestimmen, mit dem ein quadratisches Fehlermaß ε^2 minimiert werden kann. Die Adaption erfolgt in Richtung des negativen Gradienten des quadratischen Fehlers. Es kann gezeigt werden, dass die Updategleichung für das zu bestimmende Filter

$$\mathbf{w}_{n+1} = \mathbf{w}_n + \mu\,\epsilon\,\mathbf{x}[k] \tag{32}$$

lautet, mit Filter **w** der Länge L, Schrittweite μ, Fehler ε und Signalvektor **x**, der die letzten L Abtastewerte des Signals bis zum Abtastwert k enthält.

Wird der Updateterm zusätzlich auf die Leistung des Signals **x** normiert, erhält man die Updategleichung des NLMS Algorithmus [Hay96]

$$\mathbf{w}_{n+1}[k] = \mathbf{w}_n[k] + \frac{\mu\,\epsilon\,\mathbf{x}[k]}{\|\mathbf{x}\|^2} \quad . \tag{33}$$

Bei dem NLMS-basierten Ansatz zur Enthallung eines Mikrofonsignals wurde als Fehlermaß ε der Ausdruck

$$\epsilon = x[k] - \mathbf{w}[k]^T \cdot \mathbf{x}[k-\kappa] \tag{34}$$

gewählt. Der Fehler, der zugleich das Ausgangssignal darstellt, ist somit der Teil des Eingangssignals, der sich nicht aus der Prädiktion mit Hilfe des Signals berechnen lässt, das bereits vor mehr als κ Abtastwerten aufgenommen wurde. Da insbesondere Hall sehr lange Abhängigkeiten dem Signal hinzufügt, entspricht die Minimierung des Fehlermaßes somit einer Beseitigung des Halls. Die Verzögerung von κ Abtastwerten ist jedoch nötig, da sonst zusätzlich zu den Korrelationen, die der Hall verursacht, die der menschlichen Sprache ebenfalls entfernt werden (Schutzbereich, vgl. Algorithmus von Gillespie und Atlas).

6.3.4.2 Implementierung

Es wurde ein vorgegebener Offline-Algorithmus verwendet. Das Signalflussdiagramm des Programms ist in Abbildung 22 dargestellt.

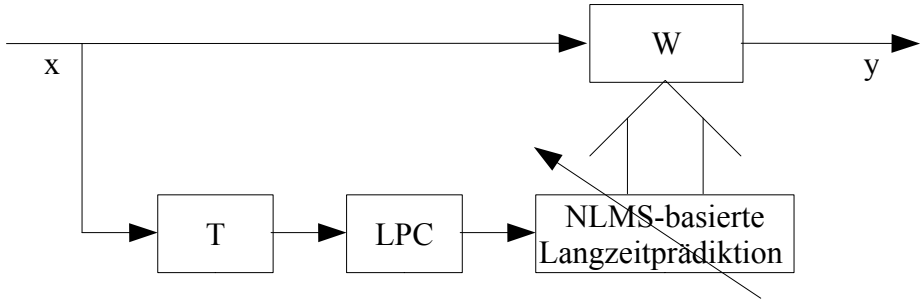

Abbildung 22: Signalflussdiagramm des NLMS-basierten Algorithmus

Um die natürliche Korrelation der Sprache nicht zu zerstören, wurde vor der Filterberechnung diese mit Hilfe linearer Prädiktion entfernt. Auf diese Weise sieht der Algorithmus im Wesentlichen nur die Korrelation, die zusätzlich durch Hall dem Signal hinzugefügt wurde.

Für die Verzögerung κ wurde ein Wert von 512 gewählt, was bei einer Abtastrate von 20.000 Hz 26 ms entspricht. Bei den Untersuchungen des Einflusses von Hall auf die WER eines Spracherkennungssystem wurde ungefähr dieser Wert als Grenze für den dramatischen Abfall der WER ermittelt.

6.3.4.3 Versuchsergebnisse

Auch hier erfolgte die Versuchsdurchführung analog zu den vorangegangenen.

Durch den NLMS-basierten Algorithmus lassen sich leichte Gewinne bei der WER von gut einem Prozentpunkt erzielen. Die WER stieg im Büroraum von 79,3% auf 80,6%.

Der Systemabstand der Gesamtimpulsantwort konnte durch diesen Algorithmus ebenfalls reduziert werden. Im Büroraum konnte er von 51,6 auf 46,7 gesenkt werden.

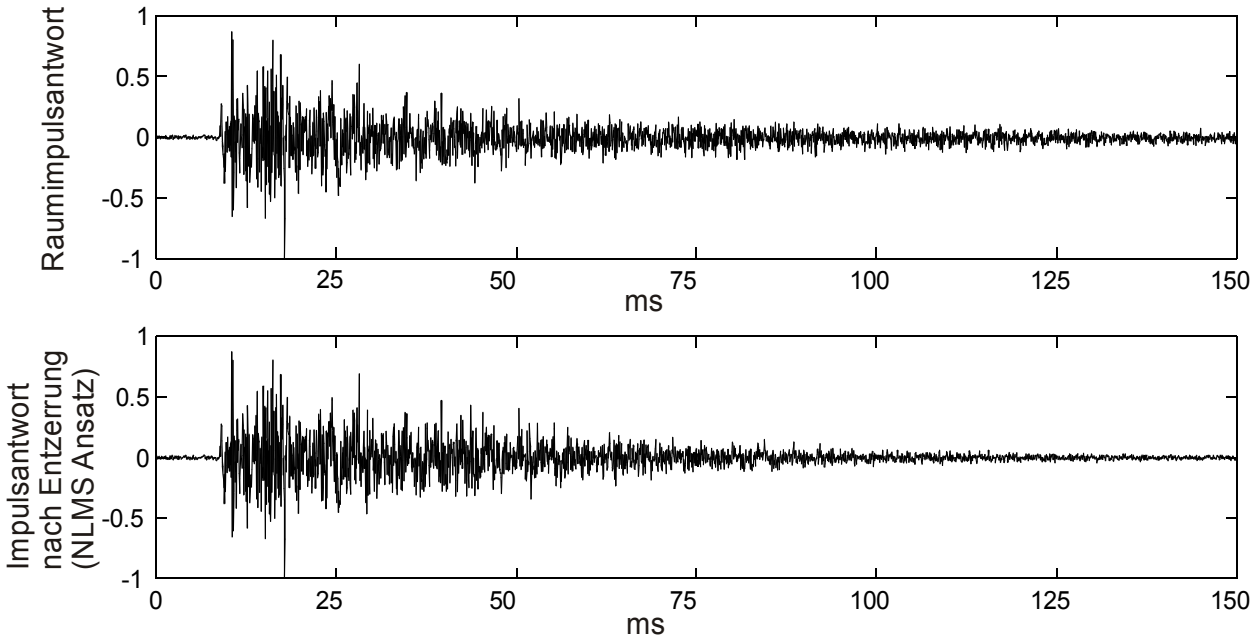

Abbildung 23: Änderung der Impulsantwort eines Büroraums durch NLMS-basierten Ansatz, Länge des Rekonstruktionsfilters 1024, Verzögerung 512

Vergleicht man die ursprüngliche Raumimpulsantwort mit der Gesamtimpulsantwort zwischen Sprecher und Systemausgang, so zeigt sich, dass besonders im hinteren Bereich die Echos reduziert werden konnten. Die Nachhallzeit wird allerdings nicht besonders stark reduziert. Im Büroraum betrug die der Gesamtimpulsantwort immer noch 130 ms. D.h., es konnten gegenüber der ursprünglichen Nachhallzeit von 150 ms nur 20 ms eingespart werden. Ein Grund für diese geringe Verbesserung ist vermutlich darin zu sehen, dass durch das Rekonstruktionsfilter die Länge der Gesamtimpulsantwort um seine Länge größer wird und diese zusätzlich hinzugefügten Anteile ebenfalls bei der Berechnung der Nachhallzeit Einfluss haben.

Im hallarmen Raum konnten zwischen der Raumimpulsantwort und der Gesamtimpulsantwort keine wesentlichen Unterschiede festgestellt werden. Der Systemabstand wurde von 7,3 auf 7,0 reduziert, die WER konnte um 0.01 Prozentpunkt auf 98,19% gesteigert werden. Die Nachhallzeit wies keinerlei Unterschiede auf. In diesem Fall brachte der NLMS-basierte Ansatz keinen Gewinn.

Im Unterschied zu den anderen beiden Verfahren, die die Statistik 2. Ordnung ausnutzen,

zeichnet sich dieser Ansatz durch eine hohe Stabilität aus. Probleme mit anwachsenden periodischen Komponenten gab es keine.

6.3.5 Ergänzende Betrachtungen

Die drei soeben beschriebenen Algorithmen basieren, wie gezeigt, alle auf dem gleichen Grundprinzip: Sie versuchen durch Gradientenabstieg eine Kostenfunktion zu minimieren. Deshalb drängt sich die Frage auf, in welcher Beziehung diese drei Algorithmen zu einander stehen.

Bevor der Updateterm des BSS-basierten Algorithmus vereinfacht wurde, wies er die Form

$$\frac{\partial \mathrm{I}}{\partial \mathbf{W}^*} = 2 \sum_{\mathrm{i}=0}^{\mathrm{m}} \beta(\mathrm{i},\mathrm{m}) \mathbf{R}_{\mathbf{xy}} \{ \mathrm{bdiag}^{-1} \mathbf{R}_{\mathbf{yy}} - \mathbf{R}_{\mathbf{yy}}^{-1} \} \tag{27}$$

auf. Mit Elementarumformungen lässt er sich in den Ausdruck

$$\frac{\partial \mathrm{I}}{\partial \mathbf{W}^*} = 2 \sum_{\mathrm{i}=0}^{\mathrm{m}} \beta(\mathrm{i},\mathrm{m}) \mathbf{R}_{\mathbf{xy}} \mathbf{R}_{\mathbf{yy}}^{-1} \{ \mathbf{R}_{\mathbf{yy}} - \mathrm{bdiag} \mathbf{R}_{\mathbf{yy}} \} \mathrm{bdiag}^{-1} \mathbf{R}_{\mathbf{yy}} \tag{35}$$

überführen. Bei der Entwicklung des BSS Algorithmus wurde vom Ansatz mit Frobenius Norm als Kostenfunktion herkommend die Normierung mit den Termen $\mathbf{R}_{\mathbf{yy}}^{-1}$ und $\mathrm{bdiag}^{-1} \mathbf{R}_{\mathbf{yy}}$ eingeführt, um Stabilitätsprobleme zu umgehen. Lässt man diese Normierungen allerdings weg, erhält man den Ausdruck

$$\frac{\partial \mathrm{I}}{\partial \mathbf{W}^*} = 2 \sum_{\mathrm{i}=0}^{\mathrm{m}} \beta(\mathrm{i},\mathrm{m}) \mathbf{R}_{\mathbf{xy}} \{ \mathbf{R}_{\mathbf{yy}} - \mathrm{bdiag} \mathbf{R}_{\mathbf{yy}} \} \quad . \tag{36}$$

Vergleicht man diesen Term mit den Updateterm des Ansatzes von Gillespie und Atlas (Gleichung 22) stellt man fest, dass sie bis auf die Tatsache gleich sind, dass in Gleichung 36 Korrelationsmatrizen und in Gleichung 22 die Auto- bzw. Kreuzkorrelierte Verwendung finden. Somit könnte erklärt werden, warum der Ansatz von Gillespie und Atlas ebenfalls zur Instabilität neigt.

Nun wäre noch der Zusammenhang zwischen dem Updateterm des NLMS-basierten Ansatzes (Gleichung 33) und dem des BSS-basierten Ansatzes zu klären.

Da ε das Ausgangssignal des Rekonstruktionssystems ist und **x** einen Vektor von

Abtastwerten des Eingangssignals darstellt, kann der Ausdruck $\varepsilon\mathbf{x}$ als eine Näherung der Kreuzkorrelierten zwischen Eingangs- und Ausgangssignel R_{xy} interpretiert werden. Die Normierung auf die Leistung des Eingangssignals $\left|(\mathbf{x})^2\right|$ wiederum kann als grobe Näherung des Ausdruckes $\{\text{bdiag}^{-1}\mathbf{R}_{yy} - \mathbf{R}_{yy}^{-1}\}$ betrachtet werden, da dieser Ausdruck ebenfalls eine Normierung allerdings auf die Ausgangsleistung darstellt. Durch die soeben durchgeführten Näherungen und Interpretationen kann die Gleichung 36 ebenfalls als Näherung der Updategleichung des BSS-basierten Ansatzes ohne natürlichen Gradienten (Gleichung 27) angesehen werden. Der wesentliche Unterschied zum Ansatz von Gillespie und Atlas besteht in der Normierung auf die Leistung des Eingangssignals, wodurch sich die größere Stabilität dieses Algorithmus erklären ließe.

Ein elementarer Unterschied zwischen dem Ansatz von Gillespie und Atlas und dem NLMS-basierten Ansatz auf der einen Seite und dem BSS-basierten Ansatz auf der anderen stellt die Verwendung des natürlichen Gradienten dar, durch den eine Entkopplung der Eingänge von der Berechnung der Filterkoeffizienten erwirkt wird. Bei BSS bringt dieser Schritt eine verbesserte Konvergenz mit sich, weshalb seine Verwendung dort äußerst wichtig ist. Bei dem vorliegenden Problem der Enthallung hingegen bewirkt er, wie bereits in 6.3.3.2 gezeigt wurde, dass die Filter der einzelnen Mikrofonsignale nicht für das jeweilige Signal getrennt optimiert werden, sondern nur ein gemeinsames optimales Filter für alle Mikrofonsignale gesucht wird. Jedoch wurde beim MINT Theorem [Miy88] gezeigt, dass gerade durch für das einzelne Mikrofonsignal optimierte Filter eine vollständige Enthallung – zumindest theoretisch – möglich ist.

7 Vergleich

Der Vergleich der betrachteten Verfahren untereinander liefert, dass sie sich in Ihrer Leistungsfähigkeit kaum unterscheiden. Während die meisten auf dem Delay-and-Sum-Beamformer aufbauen, versucht nur der Algorithmus von Gillespie und Atlas, an das einzelne Mikrofonsignal angepasste Filter zu finden, mit dem ein globales Optimum erreicht wird. Der BSS-basierte Ansatz, der NLMS-basierte Ansatz und der Cepstrum-basierte Ansatz hingegen stellen einkanalige Lösungsvorschläge vor, bei denen Mehrkanaligkeit nur durch Beamformer erreicht werden kann.

Eine direkte Ausnutzung der Kanaldiversität findet bei diesen Ansätzen also nicht statt. Dies spiegelt sich auch in den ernüchternden Gewinnen bei der WER wider (Vgl. Abbildung 24). So konnte mit dem Testsystem im besten Fall 2 Prozentpunkte Gewinn ermittelt werden, obwohl durch den Hall ein Verlust von bis zu 20 Prozentpunkten verursacht worden ist, den es galt, wieder gut zu machen.

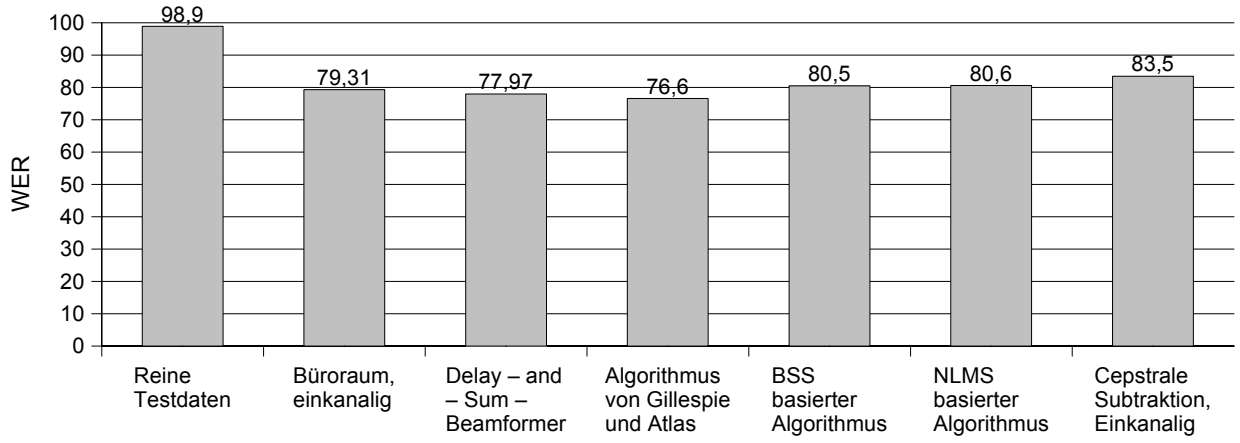

Abbildung 24: WER der einzelnen Algorithmen im Vergleich mit reinen Testdaten und durch Büroraum verhallte Testdaten

Zu beobachten ist jedoch, dass der Cepstrum-basierte Ansatz – obwohl er implementierungsbedingt in den Untersuchungen benachteiligt war – die besten Ergebnisse liefert und noch das größte Potential für weitere Gewinne besitzt.

Allerdings könnten die adaptiven Verfahren basierend auf Prädiktion oder BSS in sich zeitlich veränderlichen Umgebungen deutliche Vorteile gegenüber dem Cepstrum-basierten Verfahren besitzen. Aufgrund der langen Fensterlängen kann der Cepstrum-basierte Algorithmus nur sehr langsam den sich ändernden Raumimpulsantworten folgen.

wohingegen bei BSS gezeigt werden konnte, dass auch in sich relativ schnell veränderlicher Umgebung sehr gute Ergebnisse erzielt werden können.

Der ein- und der zweikanalige BSS-basierte Ansatz kommen dem Cepstrum-basierten Ansatz sehr nahe, wenn die richtigen Updateterme verwendet werden. Das schlechtere Abschneiden des zweikanaligen Ansatz im Vergleich, lässt sich auf den Verlust zurückführen, der bei der Addition der beiden Mikrofonsignale durch den Beamformer erlitten wurde.

Auch noch im Gewinnbereich liegt der NLMS Ansatz. Er zeichnet sich jedoch gegenüber den eben erwähnten Verfahren durch eine deutlich geringere Laufzeit aus.

Verluste hingegen erzielten unerwarteter Weise der Delay-and-Sum-Beamformer und der Algorithmus von Gillespie und Atlas. Bei beiden Verfahren ist aus Veröffentlichungen bekannt, dass sie durchaus in der Lage sind, das Signal-zu-Hall-Verhältnis zu verbessern. Worauf diese Verluste zurückzuführen sind, konnte in dieser Arbeit nicht geklärt werden.

Da aber die besseren Algorithmen zu mindest auf dem Delay-and-Sum-Beamformer bei der Verarbeitung mehrere Mikrofonsignale aufbauen, und ihre Fähigkeit zur Enthallung somit maßgeblich von diesem abhängen, wird sich nichts Wesentliches an dieser soeben vorgestellten Reihenfolge ändern.

Auf einen Vergleich der Maße Systemabstand und Nachhallzeit wird verzichtet, da es sich erwiesen hat, dass diese äußerst unzuverlässige Aussagen über die Güte der Algorithmen liefern und sich somit als nicht aussagekräftig erwiesen haben.

Zieht man zu den geringen Gewinnen die langen Laufzeiten der Algorithmen in Betracht (Abbildung 25), die zur Enthallung von 10 Sekunden Sprache nötig waren, stellt sich zunächst die Frage, ob sich dieser Aufwand in kommerziellen Systemen lohnt.

Wie bereits beschrieben, wurden die Untersuchungen mit einem Zahlenerkenner durchgeführt. Die Gewinne könnten bei Spracherkennungssystemen für kontinuierliche Sprache höher ausfallen, da bei diesen bereits geringe Verbesserungen in der Erkennung der Wortuntereinheiten wegen der zur Verfügung stehenden Grammatik zu deutlichen Verbesserung der Worterkennung führen. In Kürze dürfte auch die Rechenleistung der zur Verfügung stehenden Computer soweit angestiegen sein, dass dabei auch die großen Rechenzeiten nur noch eine untergeordnete Rolle spielen. Zudem ist eine Verringerung

der Rechenzeiten durch effiziente Implementierungen und online Algorithmen zu erwarten.

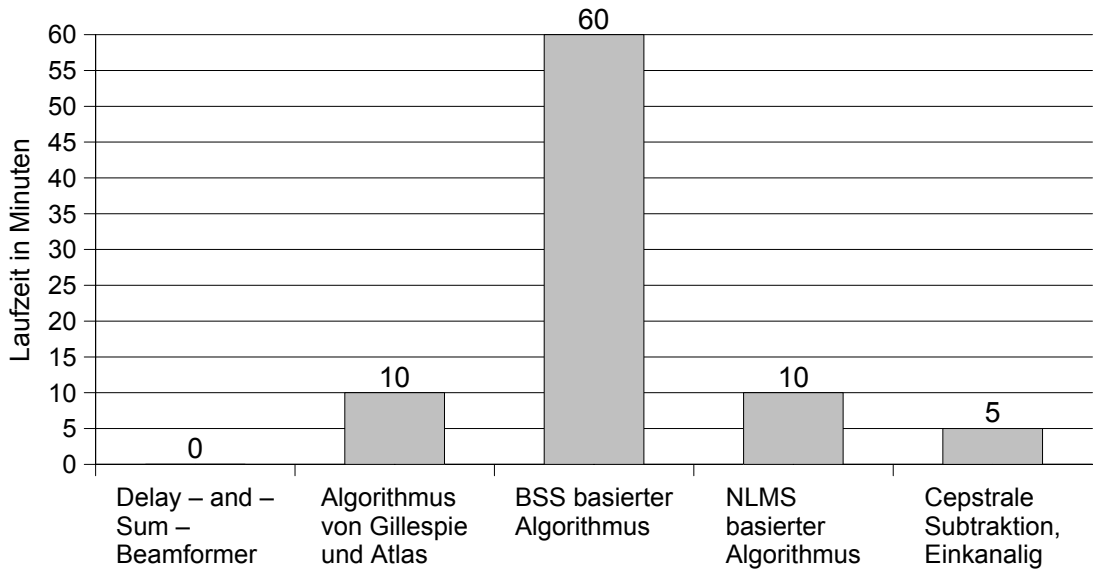

Abbildung 25: Vergleich der Rechenzeiten der einzelnen Algorithmen, um 10 Sekunden Sprache zu enthallen

Insbesondere lassen sich für alle der drei untersuchten adaptiven Verfahren effiziente Frequenzbereichsversionen unter Ausnutzung der FFT [Buc03] angeben, die bereits heute schon für zahlreiche Konfigurationen für den Echtzeitbetrieb auf handelsüblichen PC-Plattformen geeignet sind.

8 Ausblick

Hauptaugenmerk dieser Arbeit war es, die Möglichkeiten der Enthallung zu studieren und Aussagen darüber zu treffen, ob und in wie weit eine Enthallung mit dem BSS-basierten Ansatz möglich ist und wie seine Leistung mit anderen, bereits erprobten Algorithmen konkurrieren kann. In weiteren Untersuchungen sollte bei dem BSS-basierten Ansatz überprüft werden, welche Verbesserungen dadurch erzielt werden können, wenn der natürliche Gradient bei der Berechnung der Filterkoeffizienten weggelassen wird.

In den hier durchgeführten Untersuchungen wurden Störgeräusche durch weitere Quellen jedweder Art nicht zugelassen. Auch änderte sich die Raumimpulsantwort über die Zeit nicht, was z.B. der Fall bei einer bewegten Quelle wäre. Einflüsse dieser Art bedürfen ebenfalls einer weiteren Untersuchung.

Als gute Alternative für die Enthallung zu dem BSS-basierten Ansatz hat sich der Cepstrum-basierte Ansatz erwiesen. Obwohl dieser Ansatz bereits im Jahre 1969 vorgeschlagen wurde und das Cepstrum bei vielfältigen Problemen Anwendung findet, stellte es sich bei dieser Untersuchung heraus, dass sehr wenig über die Rekonstruktion eines Sprachsignals aus dem Cepstrum bekannt ist. Auch dieses Thema sollte in weiteren Arbeiten näher beleuchtet werden, zumal es auch in anderen Anwendungen (z.B. Lokalisierung eines Sprechers) sich als interessant erweisen könnte.

9 Nachwort

Ich bedanke mich bei den Mitarbeitern des Lehrstuhls Multimediakommunikation und Signalverarbeitung der Universität Erlangen-Nürnberg namentlich Dipl.-Ing Herbert Buchner, Dipl.-Ing Robert Aichner und Dipl.-Ing Wolfgang Herbordt für die mir gewährte Unterstützung, die jederzeit Zeit für mich hatten, wenn ich mich wieder einmal unangemeldet bei ihnen einfand.

Ich danke Herrn Prof. Kellermann für die wertvollen Hinweise, Anregungen und Kommentare während der Ausführung dieser Arbeit.

Des weiteren danke ich Herrn Gillespie für die entgegengebrachten Hinweise seinen Algorithmus betreffend.

10 Literaturverzeichnis

[Buc03] Buchner, H., Aichner, R., Kellermann W.: A Generalization Of A Class Of Blind Source Seperation Algorithms For Convolutive Mixtures, International Symposium on Independent Component Analysis and Blind Source Seperation(ICA), Nara, Japan, 2003

[Gil03] Gillespie, B. W. und Atlas, L. E.: Strategies For Improving Audible Quality And Speech Recognition Accuracy Of Reverberant Speech, International Conference on Acoustics, Speech and Signal Processing(ICASSP), Hongkong, China, 2003

[Gir97] Girod, B., Rabenstein, R., Stenger, A.: Einführung in die Systemtheorie, Teubner, Stuttgart 1997

[Hän01] Hänsler, E.: Statistische Signale-Grundlagen und Anwendungen. Springer Verlag, Berlin 2001

[Hay96] Haykin,S.: Adaptive Filter Theory, 3. Auflage, Prentice Hall, Englewood Cliffs, NJ., 1996

[HTK03] HTK Web-Site, htk.eng.cam.ac.uk

[Kor68] Korn, G., Korn, T.: Mathematical Handbook for scientists and engineers. McGraw Hill Book Company, New York 1968

[Miy88] Miyoshi, M. und Kaneda, Y.:Inverse Filtering of Room Acoustics. IEEE Transactions on Acoustics, Speech and Signal Processing, Vol 36, No. 2, Feb. 1988

[Nie02] Niemann, H.: Folien zur Vorlesung Mustererkennung 1, Erlangen 2002

[Nie03] Niemann, H.: Folien zur Vorlesung Mustererkennung 2, Erlangen 2003

[Sel03] Seltzer, M. L. und Raj, B.: Calibration of Microfon Arrays for Imporved Speech Recognition, www-2.cs.cmu.edu/~mseltzer/papers

[Tri79] Tribolet, J. M.: Seismic Applications of Homomorphic Signal Processing, Prentice-Hall, Englewood Cliffs, 1979

[Var98] Vary, P., Heute, U., Hess, W.:Digitale Sprachsignalverarbeitung, Teubner, Stuttgart 1998