Fraunhofer IDMT

CARL VON OSSIETZKY universität OLDENBURG

Analyse von kombinierter Grundfrequenz- und Richtungsschätzung bei simultanen Sprechern

Masterarbeit
im Studiengang
Hörtechnik und Audiologie

vorgelegt von

Stephan Gerlach

Oldenburg, 8. März 2012

Analyse von kombinierter Grundfrequenz- und Richtungsschätzung bei simultanen Sprechern

Carl von Ossietzky Universität Oldenburg
Standort Wechloy
Signalverarbeitung, Fakultät V
Carl-von-Ossietzky-Str. 9 - 11
D - 26129 Oldenburg

Fraunhofer-Institut für Digitale Medientechnologie IDMT
Projektgruppe Hör-, Sprach- und Audiotechnologie
Haus des Hörens
Marie-Curie-Straße 2
D - 26129 Oldenburg

Erstgutachter:	Prof. Dr. ir. Simon Doclo
Zweitgutachter:	Prof. Dr.-Ing. Jörg Bitzer
Betreuer:	Dipl.-Ing. Stefan Goetze

Kurzfassung

Der Zweck der kombinierten Grundfrequenz- und Richtungsschätzung ist es, anhand von akustischen Informationen der Umgebung eine Aussage über sich darin befindliche Sprecher zu machen (auf Grundlage der Aufenthaltsorte und Stimmgrundfrequenzen). Ziel dieser Arbeit ist die Analyse von Algorithmen, welche aus zwei oder mehreren Mikrofonsignalen simultan eine Grundfrequenz- und Richtungsschätzung ermittlen. Dabei werden sowohl aus der Literatur bekannte Verfahren (PoPi-Algorithmus) untersucht als auch eigene neue Erweiterungen behandelt. Aufbauend auf diesen Schätzverfahren wird zur Quellenverfolgung ein Partikel-Filter mit nachfolgender selbst entworfener Sperr-Filterung präsentiert. Ein Ergebnis dieser Arbeit ist ein Algorithmus der bei moderaten akustischen Bedingungen mehrere sich bewegende und sogar kreuzende Signalquellen detektieren, unterscheiden und verfolgen kann.

Abstract

The purpose of the joint position-pitch estimation is the localisation of active speakers based on acoustic information from the environment (direction-of-arrival and the fundamental frequency of speech). The aim of this thesis is the development of algorithms, which simultaneously estimate the pitch and direction-of-arrival using two or more microphone signals. Thereby procedures and extensions from the recent literature (PoPi algorithm) are investigated as well as own new extensions are proposed. Based on these estimation methods a particle-filter algorithm for tracking purposes with an attached self-designed notch-filtering is presented. It can be shown that the favored algorithm is able to detect, distinguish and track multiple moving and intersecting signal-sources at moderate acoustic conditions.

Inhaltsverzeichnis

Kapitel 1

Einleitung

Das Thema der vorliegenden Arbeit ist die Kombination von akustischer Richtungsschätzung sowie gleichzeitiger Grundfrequenzschätzung in Gegenwart simultaner Sprecher. Dabei versteht man unter der akustischen Richtungsschätzung im Allgemeinen die Feststellung der Einfallsrichtung einer oder mehrerer Quellen anhand der emittierten akustischen Signale. Die Grundfrequenz ist eine fundamentale Eigenschaft der menschlichen Sprache und unterscheidet sich von Mensch zu Mensch. Sie ist maßgeblich am Charakter einer individuellen Stimme beteiligt. Die Grundfrequenzschätzung stellt damit ein mögliches Kriterium zur Unterscheidung von mehreren Sprechern ebenfalls anhand ihrer ausgestrahlten akustischen Signale dar.

Die Grundfrequenz- und Richtungsschätzung von Sprache in Mehrsprechersituationen stößt in der Signalverarbeitung auf großes Interesse, um z.B. mehrere gleichzeitige Sprecher zu orten oder eine Sprechertrennung zu realisieren. Beispielhafte Anwendungsgebiete können dabei Video-Konferenzsysteme sein, die sich automatisch auf den aktiven Sprecher ausrichten, oder eine virtuelle Repräsentation der Konferenzsituation darbieten. Weiterhin können Hörgeräteträger von solchen Algorithmen profitieren. Mit Hilfe der Richtungsschätzung und Sprecherunterscheidung kann dazu ein adaptiver Beamformer gesteuert werden. Damit ist eine Fokussierung auf einen bevorzugten Sprecher in Mehrsprechersituationen möglich.

Das Ziel dieser Arbeit ist, Algorithmen zu untersuchen die eine kombinierte Auswertung einer akustischen Szene in Hinblick auf eventuelle Sprecher ermöglichen. Dabei soll die Grundfrequenz der Sprache sowie der Einfallswinkel zum Bezugspunkt ermittelt werden. Es sind zwei grundlegende Verfahren zu betrachten, die zum einen auf der zeitlichen und zum anderen auf der spektralen Repräsentation der akustischen Umgebung beruhen. Zu diesen Kernfunktionen sollen weitere, unterschiedlich motivierte Erweiterungen erläutert sowie untersucht werden. Es sollen dabei in der Literatur beschriebene sowie eigene, neue Erweiterungen verwendet werden. Im Anschluss an die Schätzalgorithmen soll ein Verfahren zur Sprecherverfolgung und Unterscheidung analysiert werden, welches robust in akustisch schwierigen Situationen (z.B. Hintergrundrauschen und Nachhall) ist und zudem für simultane Sprecher geeignet ist. Eine Evaluation soll klären welche der betrachteten Algorithmen und der darauf aufbauenden Erweiterungen das optimale Ergebnis erzielen. Die resultierende Kombination soll auf deren Robustheit für verschieden Nachhallsituationen und Störgeräusche untersucht werden. Außer-

dem soll geklärt werden, inwiefern sich bewegende Signalquellen verfolgen lassen und ob sich diese bei der Verfolgung kreuzen dürfen, da dies für die Algorithmen eine besondere Herausforderung darstellt. Zu guter Letzt soll ein Vergleich der Richtungsschätzung mit einem aus der Literatur wohl bekannten Verfahren unternommen werden.

Diese Arbeit ist wie folgt gegliedert: Im Kapitel 2 werden die physiologischen bzw. physikalischen Grundlagen für die Grundfrequenzschätzung als auch für die Richtungsschätzung erläutert. Außerdem beschreibt das Kapitel in der Literatur bekannte Verfahren um eben jene Merkmale getrennt voneinander zu ermitteln. Es werden zusätzliche Betrachtungen zur räumlichen und zeitlichen Abtastung nebst „Zeropadding“ behandelt. Kapitel 3 beinhaltet die Beschreibung der im Rahmen dieser Arbeit untersuchten Kernverfahren zur kombinierten Grundfrequenz- und Richtungsschätzung sowie die darauf aufbauenden Erweiterungen. Dabei sind die eigenen Erweiterungen, eine neue Phasentransformation, eine Multichannel Cross-Correlation (MCCC)-Integration sowie eine zusätzliche GCC-Phat Gewichtung ebenfalls Bestandteil des Kapitels. Anschließend erfolgen in Kapitel 4 die Erläuterungen zur Quellendetektion und Verfolgung mittels Partikel-Filter aufbauend auf den zuvor beschriebenen Schätzverfahren. Im zweiten Teilabschnitt wird eine in dieser Arbeit entwickelte Sperr-Filterung für das Partikel-Filter vorgestellt, die das Erkennen von simultanen Sprechern ermöglicht. In Kapitel 5 werden die beschriebenen Algorithmen evaluiert. Dabei wird im Unterschied zur bekannten Literatur neben der getrennten Erkennerrate von Grundfrequenz und Richtungsschätzung immer auch ein Augenmerk auf die kombinierte Schätzung geworfen. Es wird mit statischen Quellen aber auch mit sich bewegenden und kreuzenden Signalquellen gearbeitet und der Algorithmus mit dem wohlbekannten Generalized Cross-Correlation (GCC)-Verfahren [KC76] verglichen. Das letzte Kapitel beinhaltet eine Zusammenfassung und einen weiterführenden Ausblick.

Notation

Zur Unterscheidung von Skalaren, Vektoren und Matrizen im Zeit- und Frequenzbereich sowie zeitdiskreter und zeitkontinuierlicher Darstellung wird in dieser Arbeit die in Tabelle 1.1 zusammengefasste Notation verwendet.

	Kontinuierlicher Zeitbereich	Diskreter Zeitbereich	Kontinuierlicher Frequenzbereich	Diskreter Frequenzbereich
Skalar	$x(t), \xi(t),$	$x[k], \xi[k],$	$\mathsf{x}(e^{j\Omega}), \xi(e^{j\Omega}),$	$\mathsf{x}[n], \xi[n],$
Vektor	$\mathbf{x}(t), \boldsymbol{\xi}(t),$	$\mathbf{x}[k], \boldsymbol{\xi}[k],$	$\mathbf{x}(e^{j\Omega}), \boldsymbol{\xi}(e^{j\Omega}),$	$\mathbf{x}[n], \boldsymbol{\xi}[n],$
Matrix	$\mathbf{X}(t), \boldsymbol{\Xi}(t),$	$\mathbf{X}[k], \boldsymbol{\Xi}[k],$	$\mathbf{X}(e^{j\Omega}), \boldsymbol{\Xi}(e^{j\Omega}),$	$\mathbf{X}[n], \boldsymbol{\Xi}[n],$

Tabelle 1.1: Notation von Skalaren, Vektoren und Matrizen entnommen aus [Goe10].

Kapitel 2

Grundlagen

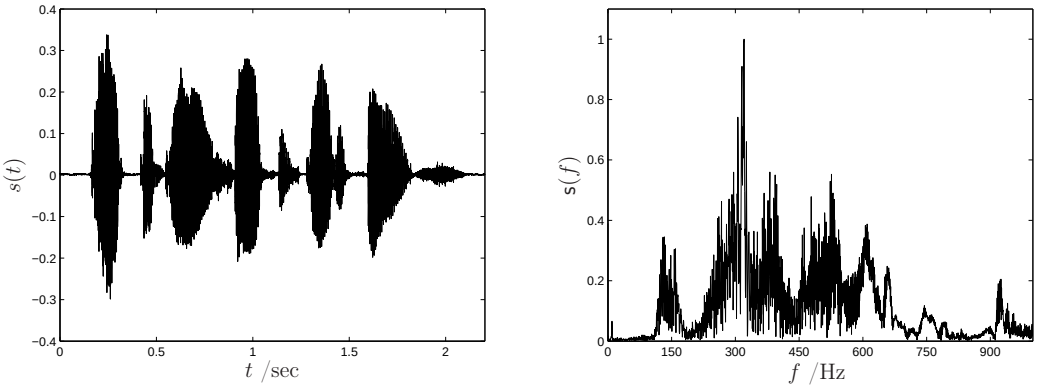
Innerhalb dieses Kapitels wird die Entstehung der Grundfrequenz der menschlichen Stimme einführend betrachtet. Zudem werden die theoretischen Grundlagen zur Entstehung einer Signaleinfallsrichtung behandelt. Im Weiteren werden weit verbreitete Algorithmen beschrieben, die sich jeweils auf einen Teilaspekt der Richtungsschätzung oder Grundfrequenzschätzung beziehen.

Im folgenden kurzen Abschnitt 2.1 werden die verwendeten Testsignale beschrieben und deren Auswahl begründet. Nachfolgend behandelt der Abschnitt 2.2 die Erzeugung von Sprache beim Menschen und sich daraus ergebende Methoden die Grundfrequenz eines Sprechers aus Mikrofonaufnahmen über die Korrelation oder das Cepstrum zu schätzen. Der zweite größere Abschnitt 2.3 dieses Kapitels befasst sich mit der Richtungsschätzung von Nutzsignalen. Ausgehend von der Beschreibung verschiedener akustischer Signalmodelle wird der Zusammenhang von Laufzeitunterschied und Einfallsrichtung erläutert, um anschließend das grundlegende Verfahren zur Richtungschätzung mittels Kreuzkorrelationsfunktion (KKF) zu erläutern. Die Abschnitte 2.3.2 und 2.3.4 geben zudem einführende Einblicke in die Problematiken der Signalabtastung und Auflösungsgenauigkeit.

2.1 Verwendete Testsignale

Um die Algorithmen zu evaluieren die in dieser Arbeit erörtert werden, sind vorrangig zwei Arten von Testsignalen verwendet worden. Da es sich um die Verarbeitung von Sprachsignalen handelt, werden zum einen Sprachaufnahmen als Testsignale verwendet. Abbildung 2.1 zeigt beispielhaft eine ungestörte Aufnahme eines Sprachsignals gesprochen von einem männlichen Sprecher im Zeit- und Frequenzbereich. Der dargestellte Satz lautet: „Wipe the grease of his dirty face.“

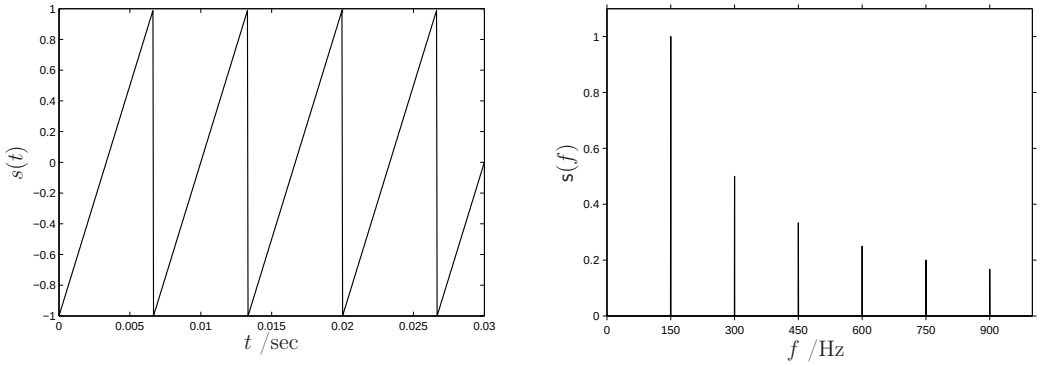
Zum anderen wird eine Sägezahnschwingung als deterministisches Signal verwendet. Die Sägezahnschwingung lässt sich aus der Fourierreihe nach Gleichung (2.1) für beliebige Frequenzen und Genauigkeiten erzeugen. Aufgrund der Ähnlichkeit des Sägezahnsignals mit der Grundfrequenz (Abschnitt 2.2.2) eines Sprachsignals eignet sich dies besonders gut als erstes Evaluationsignal. Abbildung 2.2 zeigt ein Sägezahnsignal mit einer Grundfrequenz $f_0 = 150\,\mathrm{Hz}$ im Zeit- und Frequenzbereich. Das Spektrum eines Sägezahnsignals besteht aus der Grundfrequenz und



(a) Sprachsignal. **(b)** Spektrum des Sprachsignals.

Abbildung 2.1: (a) Darstellung des Beispielsprachsignals „Wipe the grease of his dirty face.“ im Zeitbereich mit $f_s = 24\,\text{kHz}$ **(b)** Spektrum des Sprachsignals bis 1000 Hz um Vergleich mit anderen Testsignalen zu ermöglichen.

den ganzzahligen vielfachen harmonischer Oberwellen mit einer um 6 dB pro Oktave sinkender Amplitude.



(a) Sägezahnsignal mit 150 Hz. **(b)** Spektrum des Sägezahnsignals.

Abbildung 2.2: (a) Darstellung des Sägezahnsignls im Zeitbereich mit einer Grundfrequenz $f_0 = 150\,\text{Hz}$ bei $f_s = 24\,\text{kHz}$ **(b)** Spektrum des selben Sägezahnsignals bis zur fünften harmonischen Schwingung (bis 1000 Hz).

Die Fourierreihe für ein Sägezahnsignal lautet [BSM06]:

$$s[k] = \frac{2s_{\max}}{\pi}\left(\sin(\Omega k) - \frac{1}{2}\sin 2(\Omega k) + \frac{1}{3}\sin(3\Omega k)\ldots\right) \tag{2.1}$$

$$s[k] = \frac{2s_{\max}}{\pi}\sum_{p=1}^{\infty}(-1)^{p-1}\frac{\sin(p\Omega k)}{p}. \tag{2.2}$$

mit der normierten Kreisfrequenz $\Omega = \frac{2\pi f}{f_0}$ und der Spitzenamplitude $s_{\max}$.

2.2 Grundlagen zur Spracherzeugung und Grundfrequenzschätzung

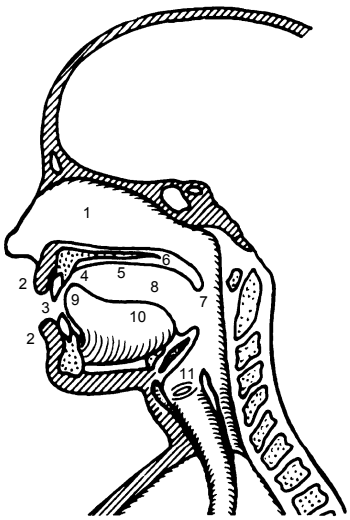
Parallel zur akustischen Richtungsschätzung des Schallereignisses (Sprache) wird in dieser Arbeit auch dessen Charakteristik analysiert. Es soll nicht Ziel sein, Personen anhand ihrer Stimme eindeutig zu identifizieren, wie es bei der Sprecheridentifikation in z.B. [PK08] der Fall ist. Vielmehr soll die Person, anhand ihrer Stimme, über einen längeren Zeitraum im Beisein mehrerer simultaner Sprecher auch bei Positionsänderungen bestimmbar sein. Da hier die Sprecherlokalisation von Interesse ist, kann als Schallsignal immer von Sprache ausgegangen werden. Im folgenden Abschnitt 2.2.1 soll daher die Spracherzeugung beim Menschen näher erläutert werden. Anschließend wird in Abschnitt 2.2.2 auf die Grundfrequenz der menschlichen Stimme als besonderes Merkmal genauer Bezug genommen. Im Abschnitt 2.2.3 wird dann auf Möglichkeiten zur Schätzung der Grundfrequenz eingegangen. Dabei werden die Korrelation und das Cepstrum als Mittel zur Grundfrequenzbestimmung diskutiert.

2.2.1 Die menschliche Spracherzeugung

Die Spracherzeugung beim Menschen ist gekoppelt an ein komplexes biologisches System, bei dem eine ganze Reihe von Organen und Muskeln an der Sprachproduktion beteiligt sind. Im Wesentlichen sind dies die Lunge, Luftröhre, Kehlkopf (mit Stimmlippen), Gaumensegel, Zunge, Lippen, Schlund und Mundmuskulatur. In Abbildung 2.3 ist eine Zeichnung mit einem Querschnitt durch den Sprachapparat dargestellt [PK08]. Die Funktionsweise des Sprachapparates ist bis heute in seinen ganzen Details noch nicht bekannt [Hol07]. Im Groben kann der Sprachapparat aber in drei Funktionsgruppen unterteilt werden: Luftstrombildung, Schallproduktion und Klangformung. Gleichzeitig muss zwischen der Erzeugung von stimmhaften, stimmlosen und plosiven Sprachlauten unterschieden werden. Anhand der eben benannten Unterteilungen soll die Spracherzeugung in diesem Kapitel näher erläutert werden.

Um mit Hilfe des Sprachapparates Klänge bzw. Sprache zu erzeugen, bedarf es zu allerst eines Luftstoms, der vom Sprachapparat akustisch beinflusst werden kann. Zur Erzeugung dieses Luftstroms dient der sogenannte Windraum, welcher aus der Lunge, den Bronchien und der Luftröhre gebildet wird. Durch Zusammenziehen des Zwerchfells und der Bauchmuskulatur wird die Luft aus der Lunge durch die Luftröhre, den Rachen und Mund- sowie Nasenraum nach außen gepresst.

Auf dem Weg nach außen passiert der Luftstrom zuerst den Kehlkopf. Hier wird der Luftstrom durch die Stimmlippen in Schwingung versetzt (Schallproduktion). Nach der *myoelastischen-aerodynamischen Theorie* [VHH98] beruht die Erzeugung der Stimmlippenbewegung auf dem Bernoulli-Effekt. Dabei wirkt auf den Luftstrom, bei Durchströmen der verhältnismäßig engen Glottis (Raum zwischen den Stimmlippen), eine Bernoullikraft quer zur Strömungsrichtung auf die



1 Nasenraum (Cavum nasi)
2 Lippen (Labia)
3 Zähne (Dentes)
4 Zahndamm (Alveolen)
5 harter Gaumen (Palatum)
6 weicher Gaumen (Velum)
7 Halszäpfchen (Uvula)
8 Mundraum (Cavum oris)
9 Zungenspitze (Apex)
10 Zungenrücken (Dorsum)
11 Stimmlippen im Kehlkopf (Larynx)

Abbildung 2.3: Seitlicher Querschnitt (Sagitalschnitt) des menschlichen Sprechapparates, entnommen aus [PK08].

Stimmlippen. Diese bewirkt, zusammen mit der Rückstellkraft der Stimmbänder, ein abruptes Verschließen der Stimmbänder, wodurch der Luftstrom kurzzeitig unterbrochen wird. Durch Schließen der Stimmbänder erhöht sich der subglottale Druck, bis die Stimmbänder diesem nicht mehr standhalten können und der Vorgang von neuem beginnt. Dabei wird das quasiperiodische Unterbrechen des Luftstromes als Ton wahrgenommen und nicht das Aneinanderpressen der Stimmlippen. In diesem Prozess begründet sich die Grundfrequenzerzeugung der Stimme, mehr dazu im Abschnitt 2.2.2.

Die Stimmlippen bestehen jeweils aus einem Bündel von Muskelgewebe, wodurch je nach Anspannung und Verformung Einfluss auf die Schwingungsanregung genommen werden kann. Bei Männern sind die Stimmlippen in der Regel $1,7\,\text{cm}$ bis $2,4\,\text{cm}$ groß, bei Frauen fallen sie mit $1,3\,\text{cm}$ bis $2,0\,\text{cm}$ etwas kleiner aus [Hol07].

Sprachlaute, die mit Hilfe der schwingenden Stimmlippen entstehen, werden als stimmhafte Sprache bezeichnet (z.B. /I/gel). Durchströmt die Luft den Sprachapparat bei dauerhaft geöffneten Stimmlippen werden die Sprachlaute stimmlos (z.B. /F/isch) genannt. Dabei wird durch Reibung im Mund und Rachenraum an einer Engstelle die Luftströmung verwirbelt und es entsteht ein akustisches Rauschen, dessen Charakteristik von der erzeugenden Engstelle abhängt [VHH98]. In Abbildung 2.4 ist das Spektrogramm des Wortes „Seewespe“ dargestellt, in dem stimmhafte von stimmlosen Sprachlauten gut unterscheidbar sind. Das Spektrogramm trägt die Sprachenergie in Abhängigkeit der Frequenzen über die Zeit auf. In Zeitabschnitten in denen stimmlose Sprachlaute liegen (z.B. Sekunde $0,4$ bis $0,5$), ist keine besondere Struktur über die Frequenzen erkennbar. Dies verdeutlicht den rauschartigen Charakter der stimmlosen Anregung. Bei stimmhafter Anregung (z.B. Sekunde $0,1$ bis $0,2$) ist hingegen eine harmonische Struktur auszumachen.
Im supraglottalen Bereich, dem sogenannten Vokaltrakt oberhalb der Glottis, er-

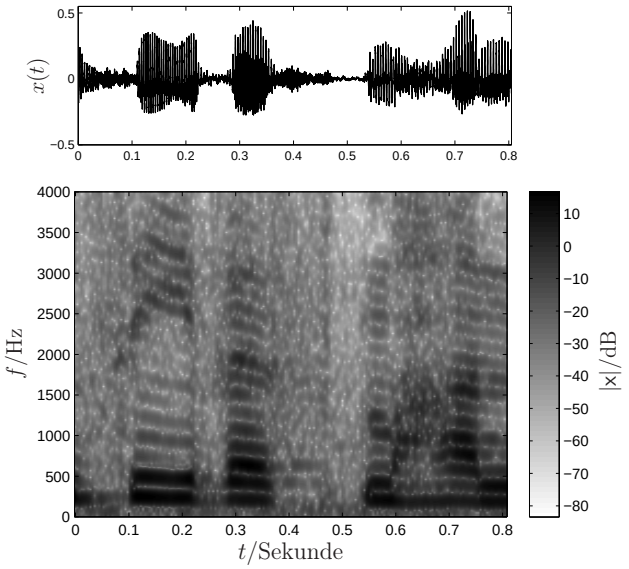


Abbildung 2.4: Spektrogramm des gesprochenen Wortes „Seewespe“. Darüber ist das äquivalente Zeitsignal aufgetragen. Die Abtastrate liegt bei 8 kHz. Die Kurzzeitspektren werden aus 16 ms langen Signalblöcken erzeugt, bei einer Überlappung von 75 %. Die stimmhaften Abschnitte sind gut von den stimmlosen Sprachlauten zu unterscheiden. In den stimmhaften Abschnitten ist eine harmonische Struktur, verursacht durch die Stimmlippenschwingungen, auszumachen.

folgt die Klangformung des Sprachapparates. Dabei dienen die veränderlichen Artikulatoren wie Zunge, Lippe, Gaumensegel und Unterkiefer als akustische Filter. Der Vokaltrakt kann grob als akustische Röhre, mit veränderbarer Form, angesehen werden. Durch Veränderung der Artikulatoren wird die Übertragungsfunktion des Vokaltraktes und damit auch dessen Resonanzfrequenz verändert. Durch diese Bearbeitung des Luftstroms entstehen die vornehmlich (stimmhaften) Sprachlaute, gekennzeichnet durch bestimmte Formantenstrukturen (siehe Abschnitt 2.2.2). Die Verformungen der Artikulation führen bei stimmloser Sprachanregung zu den bereits erwähnten rauscherzeugenden Engstellen im Vokaltrakt [PK08]. Plosive Sprachlaute (z.B. /P/ark) werden hingegen durch Bildung eines erhöhten Luftdrucks im Mundraum und anschließenden plötzlichen Öffnen der Lippen verursacht.

Eine weitere Unterteilung der Sprachlaute ist durch das Entkoppeln bzw. Ankoppeln der Nasenhöhlen an die Mundhöhle möglich. Indem der Gaumensegel die Öffnung zu Nasenhöhle frei gibt, wird das Volumen der Nasenhöhle an das Volumen der Mundhöhle angekoppelt und es entstehen nasale Sprachlaute (z.B. /N/ase). Umgekehrt werden bei Entkopplung der Nasenhöhle die Laute als nichtnasal bezeichnet.

2.2.2 Die Grundfrequenz und deren harmonische Oberwellen

Neben den Formanten, der Lautdauer und der Intensität ist die Grundfrequenz[1] eine bedeutende Charakteristik von menschlicher Sprache [PK08]. Sie wird von den Stimmbändern während des Sprechens von stimmhaften Sprachlauten erzeugt. Die stimmhaften Sprachlaute weisen eine quasi-periodische Struktur mit der Grundfrequenz und deren harmonischen Oberwellen auf. Die Schwingungsrate der Stimmbänder kann 50 Perioden pro Sekunde bei Männern und bis zu 400 Perioden pro Sekunde bei Kindern betragen. Somit kann die Grundfrequenz f_0 als Inverse der Periodendauer T_0 der Stimmbandschwingung definiert werden [VHH98]:

$$f_0 = \frac{1}{T_0}. \tag{2.3}$$

Ein Problem bei der Grundfrequenzbestimmung ist deren Veränderung über der Zeit. Der zeitliche Mittelwert der Grundfrequenz wird vom Menschen als Stimmlage gehört, wohingegen die Variation über die Zeit als Sprachmelodie wahrgenommen wird. Bei einer Grundfrequenzbestimmung über einen längeren Zeitraum und Zuordnung zu einer Person muss dieser Sachverhalt berücksichtigt werden. Die Grundfrequenz ist mehr als Tonkomplex denn als Reinton zu verstehen. Verursacht durch deren Entstehungsart beschreibt das quasi-periodische Volumenflusssignal des Luftstromes oberhalb der Stimmlippen, welches alleinig die Grundfrequenz und deren Oberwellen enthält, eher ein Sägezahnsignal. Zusätzlich sind die harmonischen Oberwellen auf Höhe der Lippen bereits durch etwaige Resonanzen des Vokaltraktes beeinflusst.

Die Oberwellen der Grundfrequenz weisen bei einem Kurzzeitspektrum von Sprache ausgeprägte lokale Maxima auf (Abbildung 2.5). Als Ursache dafür sind genau jene beschriebenen Resonanzen des Vokaltraktes zu nennen. Die Maxima werden Formanten genannt und charakterisieren bestimmte Laute. Ein Formant ist durch dessen Frequenz, Amplitude und Bandbreite definiert. Eine Formantenstruktur kann als Ganzes einem Vokal zugeordnet werden.

2.2.3 Korrelation und Cepstrum zur Grundfrequenzschätzung

Die Grundfrequenzschätzung findet in vielen aktuellen Sprachverarbeitungsalgorithmen Verwendung, um Wortgrenzen zu finden, die Satzanalyse zu verbessern oder den Satzfokus anhand der Wortbetonung auszumachen [KWK07]. Aktuelle Verfahren verwenden die Korrelationsmethode als Strategie zur Grundfrequenzschätzung, wobei Signale auf ihre Ähnlichkeit verglichen werden [KWK07, dCK02]. In diesem Abschnitt soll die Korrelation und das Cepstrum zur Grundfrequenzschätzung erläutert werden.

[1] Im Englischen wird die Grundfrequenz (engl. fundamental frequency) auch als Pitch bezeichnet.

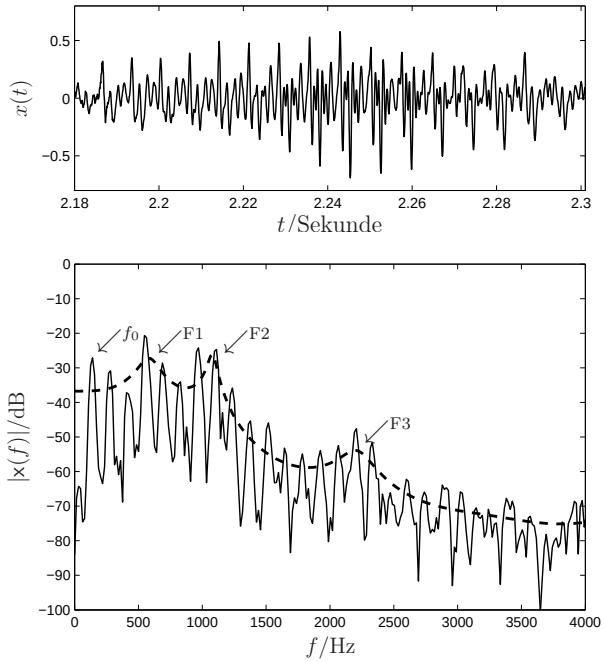


Abbildung 2.5: Ein 50 ms Kurzzeitspektrum eines stimmhaften Spachsignals. Die Grundfrequenz f_0 liegt bei 138 Hz. Die Formanten F1, F2 und F3 sind mittles einer LPC-Analyse ermittelt und eingezeichnet. Das obere Bild stellt einen Ausschnitt des zugehörigen Zeitsignals dar, in dem die Periodizität der stimmhaften Sprache zu erkennen ist.

Die Kreuzkorrelationsfunktion (KKF) zwischen zwei Signalen $x_i[k]$ und $x_l[k]$ definiert sich im Allgemeinen über den Erwartungswert

$$r_{x_i x_l}[\kappa] = \mathrm{E}\{x_i[k] x_l[k+\kappa]\}, \tag{2.4}$$

wobei das Signal $x_l[k]$ um den Wert κ verschoben wird. Aus dieser Gleichung kann die Schätzvorschrift zur Korrelationsbestimmung (Gleichung (2.5)) abgeleitet werden. Wobei die Größe N die Länge des betrachteten Signalausschnitts definiert.

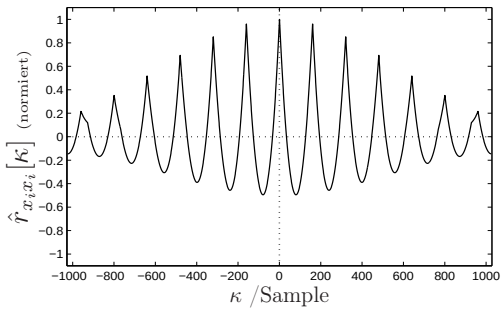
$$\hat{r}_{x_i x_l}[\kappa] = \frac{1}{N} \sum_{k=0}^{N-1} x_i[k] x_l[k+\kappa] \tag{2.5}$$

$$\hat{r}_{x_i x_l}[\kappa] = x_i[k] * x_l[-k] \tag{2.6}$$

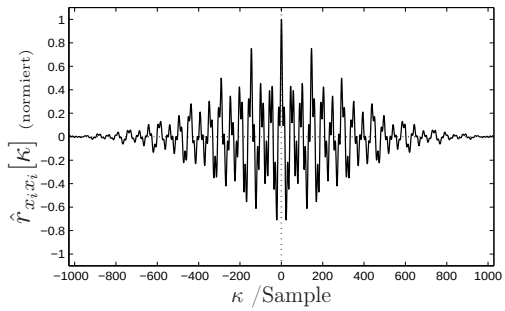
Gleichung (2.5) kann als Faltung aufgefasst werden, nur mit dem Unterschied, dass keine zeitliche Spiegelung eines der beiden Signale vorgenommen werden muss. So kann Gleichung (2.5) auch durch Gleichung (2.6) ausgedrückt werden.

Wenn die Signale $x_i[k] = x_l[k]$ identisch sind, spricht man von einer Autokorrelationsfunktion (AKF), wohingegen bei der Verwendung von zwei unterschiedlichen Signalen die Bezeichnung KKF gebraucht wird. Bei der AKF befindet sich das Maximum der Korrelationsfolge immer in dessen Urspung ($\kappa = 0$). Weitere Informationen zur Korrelationsberechnung finden sich Abschnitt 2.3.5, über die Berechnung der Korrelation mit Hilfe des Leistungsdichtespektrum (LDS).

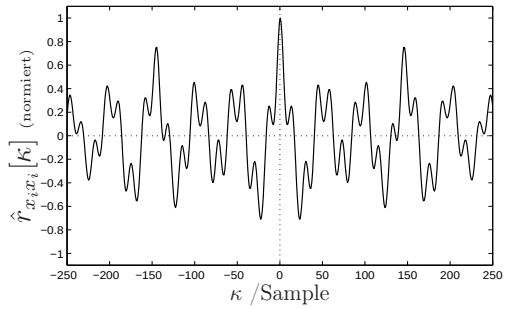
Die Grundfrequenz eines Sprachsignals äußert sich als periodisch wiederkehrende Maxima im Korrelationssignal. Hierbei entspricht der Abstand der Maxima der Periodendauer der Grundfrequenz. Bei Bestimmung der Grundfrequenz eines unverhallten und unverrauschten Sprachsignals liefert die Korrelation bereits sehr gute Ergebnisse. Wenn das zu untersuchende Sprachsignal jedoch zusätzliche Störfaktoren (Rauschen, Reflexionen) enthält, gelangt die Grundfrequenzschätzung schnell an ihre Grenzen. Störeinflusse überlagern das Korrelationssignal mit zusätzlichen Maxima, wodurch eine Erkennung der durch die Grundfrequenz bestimmten Maxima erschwert, bzw. unmöglich gemacht wird. Abbildung 2.6 zeigt drei unterschiedliche Korrelationen, die besonders ausgeprägten Maxima kodieren dabei die Grundfrequenzen.



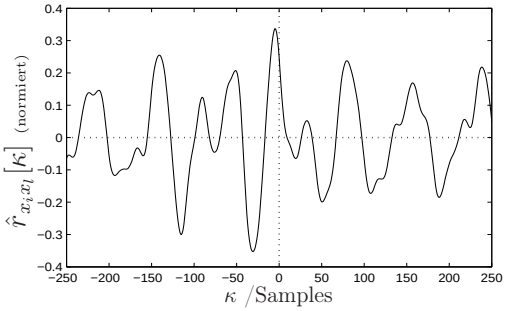
(a) Autokorrelation eines Sägezahnsignals.



(b) Autokorrelation eines Sprachsignals.



(c) Auschnitt aus Abbildung **(b)**.



(d) Autokorrelation eines verhallten Sprachsignals.

Abbildung 2.6: **(a)** Darstellung der Autokorrelation eines Sägezahnsignals mit 150 Hz und $f_s = 24\,\text{kHz}$, sowie der Autokorrelation eines Sprachsignals in unverhallter Umgebung **(b)**. Grafig **(c)** entspricht einem Ausschnitt aus **(b)** zum Vergleich mit **(d)**. Grafig **(d)** illustriert die Autokorrelation des Sprachsignals in verhallter Umgebung.

Eine weitere Methode zur Bestimmung der Grundfrequenz bietet das Cepstrum[2]. Bei dem Cepstrum handelt es sich, neben der Zeit und Spektraldarstellung eines Sprachsignals, um eine weitere Darstellungsmöglichkeit. Dabei wird das Spektrum

[2] „Cepstrum“ stammt von Begriff „Spectrum“ (deut. Spektrum) ab, dabei werden die Buchstaben der ersten Silbe vertauscht. Analog werden die Werte eines Cepstrums, abgeleitet von Frequenzen als „Quefrenzen“ bezeichnet.

eines Signals logarithmiert und anschließend wieder invers Fouriertransformiert. Durch die Rücktransformation befindet man sich streng genommen wieder im Zeitbereich. Auf Grund der vorherigen Logarithmierung wird das Resultat jedoch als Cepstrum bezeichnet. Der schematische Ablauf der Cepstrumsbestimmung ist in Abbildung 2.7 dargestellt. Bekanntermaßen wird bei der Fouriertransformation aus einer Faltung zweier Signale im Zeitbereich eine Multiplikation im Frequenzbereich. Mit der Logarithmierung der Spektren wir aus dieser Multiplikation eine Addition.

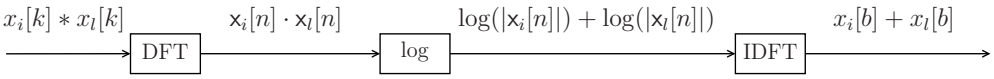


Abbildung 2.7: Schematischer Ablauf einer Cepstrumstransformation. Die Anwendung der Logarithmierung und der zweifachen DFT führt von einer Faltung zweier Signale im Zeitbereich zu einer Addition im Cepstralbereich.

Durch die Linearität der anschließenden inversen Fouriertransformation bleibt der additive Charakter der logarithmierten Spektren erhalten. Hierin ist der Vorteil des Cepstrums zu sehen. Denn die, während der Erzeugung und Übertragung des Sprachsignals, einwirkenden Systeme werden hierdurch Additiv dargestellt. Im cepstralen Bereich lässt sich somit das Anregungssignal der Sprache näherungsweise von den übrigen Einflüssen trennen. Aufgrund der Tatsache, dass die Discrete Fourier Transformation (DFT) und Inverse Discrete Fourier Transformation (IDFT) bis auf einen Normierungsfaktor und ein Vorzeichen identisch sind,

$$\text{DFT}: \quad \mathsf{x}[n] = \sum_{k=0}^{N-1} x[k] \exp^{-j2\pi nk/N}, \qquad n = 0, 1, 2, \ldots, N-1 \tag{2.7}$$

$$\text{IDFT}: \quad x[k] = \frac{1}{N} \sum_{k=0}^{N-1} \mathsf{x}[n] \exp^{j2\pi nk/N}, \qquad k = 0, 1, 2, \ldots, N-1, \tag{2.8}$$

kennzeichnen die tiefen Indizes (Quefrenzen) die Grobstruktur eins Spektrums. Die hohen Quefrenzen stellen hingegen die Feinstruktur eines Spektrums dar. Aus diesem Zusammenhang ergibt sich, dass die tiefen Quefrenzen bei einem Sprachsignal die Klangformung darstellen, wohingegen die hohen Quefrenzen im wesentlichen das Anregungssignal der Sprache spezifizieren. In Abbildung 2.8 wird solch ein Kurzzeitcepstrum eines Sprachsignals gezeigt. Der durch die Grundfrequenz erzeugte Anteil ist gut als einzelnes Maximum im Signal sichbar. Die abfallende Flanke bei kleinen Quefrenzwerten ist näherungsweise als Übertragungsfunktion der Klangformung anzusehen [VHH98, PK08, Nol67].

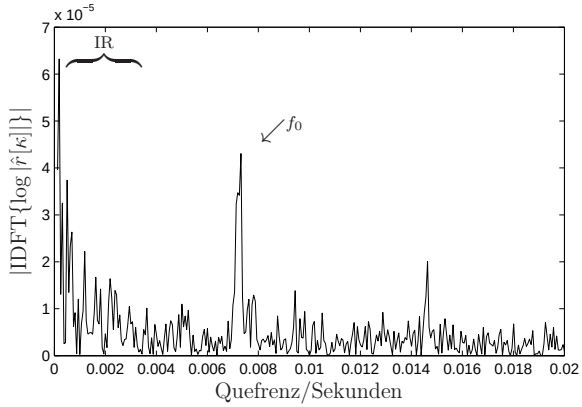


Abbildung 2.8: Reales Kurzzeitcepstrum des Signals aus Abbildung 2.5. Das Maximum bei $0,0072\,\mathrm{s}$ beschreibt die Grundfrequenz mit $f_0 = 138\,\mathrm{Hz}$. Die abfallende Struktur zu Beginn des Fensters stellt die Impulsantwort des aufgenommenen Systems dar.

2.3 Grundlagen zur Richtungsschätzung

Die Schätzung der Einfallsrichtung von akustischen Quellen ist ein Forschungsgebiet mit großer praktischen Bedeutung in vielen unterschiedlichen Anwendungsgebieten, wie z.B. in der Radar-, Sonar- und Ultraschallforschung. In der hier vorliegenden Arbeit bezieht sich die Richtungsschätzung auf die Erkennung von akustischen Quellen, im Spezifischen auf die Erkennung von Sprechern (Sprache), in der Umgebung der verwendeten Sensoren (Mikrofone). Bei der Richtungsschätzung muss zunächst eine Unterscheidung zwischen den Methoden der Time of Arrival Estimation (ToA) und Time Difference of Arrival Estimation (TDoA) gemacht werden.

Die ToA wird in diesem Zusammenhang auch als aktive Methode bezeichnet. Dabei wird ein dem Sender bekanntes Signal (z.B. ein Impuls) ausgestrahlt und die Zeit bis zur Registrierung des Echos gemessen. Bei dieser Art der Richtungsschätzung (z.B. Radar) ist nur ein Sensor notwendig. Solche Methoden finden in dieser Arbeit jedoch keine Berücksichtigung.

Bei den Strategien der TDoA werden hingegen immer mindestens zwei räumlich von einander getrennte Sensoren benötigt. Deren relative Lage zueinander muss dabei bekannt sein. Die von solch einem Sensorpaar aufgenommenen Signale können anschließend im Bezug auf ihre Laufzeitunterschiede ausgewertet werden. Die verglichenen Sensorsignale enthalten dabei keine vom System aktiv erzeugten Signale. Solch eine Art der Richtungsschätzung wird dementsprechend auch als passive Methode bezeichnet. Für ideale Freifeldbedingungen kann gezeigt werden, dass die TDoA Algorithmen eine sehr gute Schätzung der Einfallsrichtung erreichen. Eine besondere Herrausforderung ist es, die umgesetzten Algorithmen robust gegen die Störeinflüsse von Rauschquellen und Reflexionen zu machen. Rauschquellen sind in alltäglichen Situationen allgegenwertig und kön-

nen praktisch nicht gänzlich unterdrückt werden. Das Rauschen verringert das Signal-Rausch-Verhältnis (engl. Signal to Noise Ratio) (SNR) und die Schätzug der TDoA wird verschlechtert [CBH06].

Die folgenden Unterkapitel 2.3.1 bis 2.3.6 befassen sich mit den Grundlagen der Sprecherlokalisation. Um das Phänomen der Reflexionen näher zu erläutern, werden in Abschnitt 2.3.1 drei unterschiedliche Signalmodelle vorgestellt. Anhand dieser Modelle kann die Schallausbreitung in verschieden komplexen Situationen beschrieben und modelliert werden. Anschließend wird in Abschnitt 2.3.2 auf das zeitliche sowie räumliche Abtasttheorem und deren Restriktionen für die Richtungsschätzung hingewiesen. Weiterhin wird in Abschnitt 2.3.3 der Zusammenhang vom Laufzeitunterschied und Einfallsrichtung erläutert. In Abschnitt 2.3.5 wird die Kreuzkorrelation als Methode zur Feststellung des Laufzeitunterschiedes beschrieben und deren effektive Berechnung über das LDS dargestellt. Zuletzt enthält dieses Kapitel noch den Abschnitt 2.3.6 über die Generalized Cross-Correlation (GCC) Methode, als Spezialfall der Kreuzkorrelation. Es existieren noch weitere TDoA-Methoden, wie die blinde Übertragungsfunktionsbestimmung [BSH08] oder statistische Methoden, auf die an dieser Stelle der Vollständigkeit halber hingewiesen werden soll [BSH08].

2.3.1 Signalmodell

Die folgenden Signalmodelle beschreiben die Ausbreitung von einer Nutzschallquelle $s[k]$ im Raum in Bezug zu den aufnehmenden Schallsenken $x[k]$. Zusätzlich zu der Nutzschallquelle wird angenommen, dass noch eine weitere Rauschquelle $q[k]$ vorhanden ist. Dabei gehen den Modellen unterschiedliche Annahmen vorraus, die ihre Komplexität bestimmen. Die gewählte Reihenfolge entspricht der ansteigenden Komplexität der Signalmodelle.

Freie Schallausbreitung

Bei dem Modell der freien Schallausbreitung wird davon ausgegangen, dass sich die Schallwellen vollkommen ungehindert ausbreiten können. Es gibt keine störenden Einflüsse an denen die Schallwellen gebrochen oder reflektiert werden. Das Rauschsignal $q[k]$ wird, wie auch bei den folgenden Signalmodellen, als ein mittelwertfreier, zum Nutzsignal unkorrelierter Zufallsprozess angesehen, der vollkommen unkorreliert mit dem Nutzsignal $s[k]$ ist [CBH06]. Ausgehend von diesen Definitionen kann das von den Mikrofonen (M = Anzahl der Mikrofone) aufgenommene Signal als

$$x_i[k] = a_i s[k - \kappa_i] + q_i[k], \qquad i = 1, 2, \cdots, M \tag{2.9}$$

beschrieben werden. Dabei stellen die Signale an den Mikrofonen $x_i[k]$ eine um κ verzögerte und, auf Grund des räumlichen Abstandes zur Quelle, um einen Faktor a gedämpfte Version des Originalsignals dar. Diese werden nun noch zusätzlich

von dem diffusen Rauschsignal $q_i[k]$ überlagert. Solch eine Art der Schallausbreitung ist in der Realität nur unter Freifeldbedingungen oder nährungsweise in reflexionsarmen Räumen zu finden.

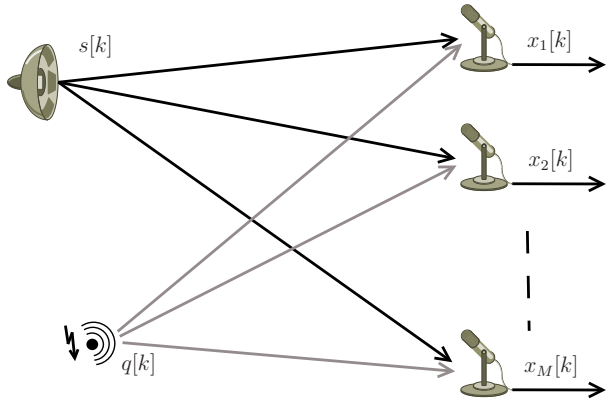


Abbildung 2.9: Signalmodell der freien Schallausbreitung mit einer Nutzschallquelle $s[k]$, einem zum Nutzsignal unkorrelierten Rauschen $q[k]$ und mehreren Sensoren (Mikrofonen) und deren Signal $x_i[k]$. Es sind nur direkte Schallpfade von der Quelle zu den Sensoren möglich.

Multipfadmodell

Das zuvor beschriebene Modell der freien Schallausbreitung berücksichtigt nur den Direktpfad des Nutzsignals zu den Mikrofonen. Im Gegensatz dazu lässt das Multipfadmodell mehrfache Ausbreitungswege des Schallsignals zu. Damit können Reflexionen des Schallfeldes an Oberflächen (z.B. Wänden oder Stühlen) beschrieben werden. Die durch die Reflexionen verursachten Signalanteile können als zusätzliche, gedämpfte und verzögerte Versionen des Originalsignals angesehen werden. Eine sich daraus ergebende mathematische Beschreibung des Mikrofonsignals sieht dann folgendermaßen aus:

$$x_i[k] = \sum_{p=1}^{P} a_{ip} s[k - \kappa_{ip}] + q_i[k] \qquad i = 1, 2, \cdots, M. \tag{2.10}$$

Das Mikrofonsignal setzt sich somit aus der Summe aller vorhandenen Signalpfade P zusammen. Diese sind gekennzeichnet durch je einen eigenen Dämpfungsfaktor a_p und einer der Laufzeit entsprechenden Verzögerung κ_p des Nutzsignals $s[k]$. Auch hier kommt das Rauschen $q_i[k]$ nochmals additiv hinzu. Das Multipfadmodell wird oft für die Signalmodellierung in großen Gewässern angewendet. In solchen Umgebungen wird davon ausgegangen, dass der Sensor nicht nur den Direktpfadanteil des Nutzsignals, sondern auch die Reflexionen vom Seeboden und Wasseroberfläche aufnimmt [Mey08, CBH06].

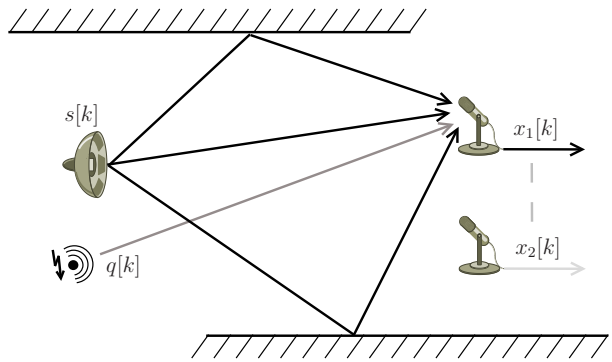


Abbildung 2.10: Darstellung des Multipfadmodells mit einer Nutzschallquelle $s[k]$, einem zum Nutzsignal unkorrelierten Rauschen $q[k]$ und mehreren Sensoren (Mikrofonen) $x_i[k]$. Es existieren mehrere, unterschiedlich lange Ausbreitungswege von der Signalquelle zu den Mirkofonen.

Nachhallmodell

In vielen alltäglichen Situationen ist es nicht möglich bzw. unpraktikabel alle Schallpfade einzeln zu simulieren. Dies ist z.B. bei der Schallausbreitung in geschlossenen Räumen der Fall (vgl. Abbildung 2.11). In solchen Umgebungen finden vielfache Reflexionen und Dämpfungen an den Raumbegrenzungen und störenden Gegenständen statt. Somit kommt es zu einer großen Anzahl sich überlagernder Schallpfade, die sich zusammen in der Raumimpulsantwort (engl. Room Impulse Response) (RIR) für jeweils eine bestimmte Quellen- und Sensorposition beschreiben lassen. Wenn die RIR bekannt ist, kann das Sensorsignal $x_i[k]$ durch Faltung des Nutzsignals $s[k]$ mit der Impulsanwort $h_i[k]$ erzeugt werden. Dieses Modell ist in Gleichung (2.11) mit zusätzlicher Addition von unkorrelierten Rauschen dargestellt [CBH06],

$$x_i[k] = h_i[k] * s[k] + q_i[k] \qquad i = 1, 2, \cdots, M. \tag{2.11}$$

In Abbildung 2.12 ist eine simulierte RIR mit einer Nachhallzeit von 500 ms dargestellt [Hab10]. Die RIR wird im Allgemeinen in drei Abschnitte aufgeteilt. Zu Beginn ist meist eine eindeutige Ausprägung zu erkennen, diese stellt den Direktpfad des Signals von der Quelle zur Senke mit dessen Verzögerungen dar (vgl. Abschnitt zur freien Schallausbreitung). Auf Grund des kürzesten Laufweges ist die Verzögerung des Direktpfades am geringsten. In den Amplituden der RIR ist die Dämpfung des Signal enthalten. Der Direktpfad besitzt oft auch die größte Amplitude. Durch konstrukive Überlagerung mehrerer Reflexionen gleicher Laufzeit kann das Maximum der RIR jedoch auch nach dem Direktpfad liegen. Anschließend sind die frühen Reflexionen des Signalpfades dargestellt (vgl. Multipfadmodell). Diese beiden Komponenten sind maßgeblich für die akustische räumliche Wahrnehmung der Menschen verantwortlich [Bla74]. Zuletzt wird der diffuse Nachhall abgebildet. Der Nachhall besteht aus sich selbst überlagernden, mehrfachen Reflexionen, bei denen eine klare Trennung der Schallpfade kaum mehr möglich ist. Bei der Synthese einer RIR wird der Bereich des Nachhalls

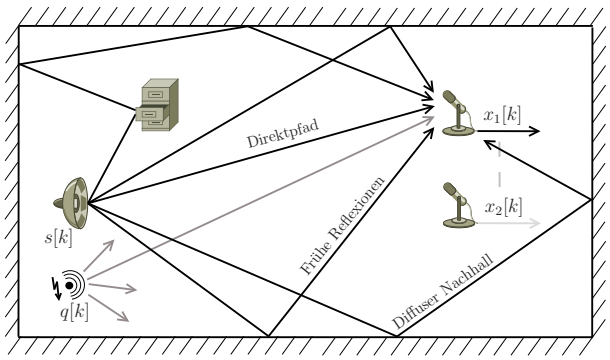


Abbildung 2.11: Exemplarische Darstellung eines Nachhallmodells mit einer Nutzschallquelle $s[k]$, einem zum Nutzsignal unkorrelierten Rauschen $q[k]$ und mehreren Sensoren (Mikrofonen) $x_i[k]$. Es existieren mehrere, unterschiedlich lange Ausbreitungswege von der Signalquelle zu den Sensoren (Mikrofonen), gekennzeichnet als Direktpfad, frühe Reflexion und diffuser Nachhall.

dementsprechend durch weißes Rauschen mit entsprechendem Amplitudenabfall modelliert. Charakteristisch für den Nachhall ist vor allen dessen Dauer, welche mit der Nachhallzeit τ_{60} (Abfall der Signalenergie um 60 dB) beschrieben wird.

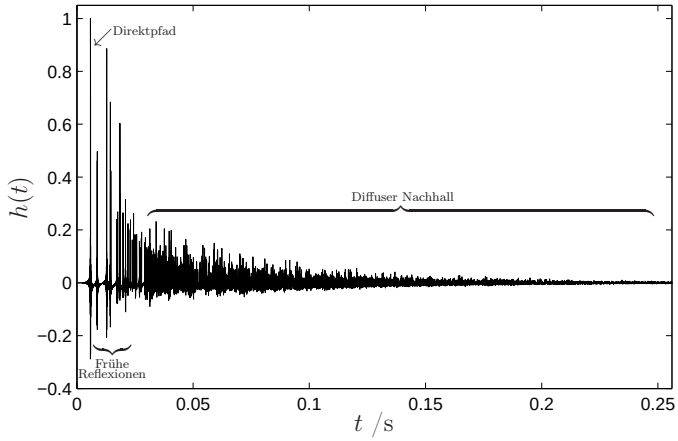


Abbildung 2.12: Simulierte Raumimpulsantwort $h[k]$ [Hab10] bei einer Nachhallzeit von 500 ms und einer Abtastfrequenz von 48 kHz. Der charakteristische Aufbau mit Direktpfadanteil, frühen Reflexionen und Nachhall ist hier gut wiederzuerkennen.

2.3.2 Zeitliches und räumliches Abtasttheorem

In dieser Arbeit sind alle Verfahren mittels wert- und zeitdiskreter Signalverarbeitung realisiert. Bei der Umwandlung von analogen Messsignalen in diskrete Signale sind gewisse Einschränkungen, besonders bei der TDoA, zu beachten. Die Wahl der konstanten Abtastfrequenz f_s, für die Abtastung eines kontinuierlichen Signals, legt den darstellbaren Frequenzbereich fest. Das Nyquist-Theorem

fast diesen Zusammenhang auf und fordert, dass die Abtastfrequenz mindestens doppelt so groß ist wie die maximal darzustellende Frequenz $f_{\max}$.

$$f_{\max} < \frac{f_s}{2} \tag{2.12}$$

Falls Gleichung (2.12) nicht eingehalten wird, können Aliasing-Effekte (Mehrdeutigkeiten) auftreten. In einer anderen Betrachtungsweise kann Aliasing auch als spektrale Überlappung des Signals im Frequenzbereich angesehen werden. Mehr Informationen zu diesem Thema ist z.B. in [KJ05] zu finden.

Die Periodendauer T kann durch die Wellenlänge λ repräsentiert werden. Diese Größen hängen über die Schallgeschwindigkeit[3] in direktem Zusammenhang miteinander,

$$\lambda = \frac{c}{f} = c \cdot T. \tag{2.13}$$

Wenn man sich die Phasenlage der Frequenz zu einem festen Zeitpunkt im Raum betrachtet, besitzt diese an zwei unterschiedlichen Orten innerhalb der Wellenlänge zwei unterschiedliche Phasen. Daher ist es möglich, über eine zeitgleiche Messung des Schallfeldes an zwei unterschiedlichen Positionen auf die Frequenz zu schließen (Vorraussetzung: sinusodiale Signale). Der Abstand zwischen den beiden Punkten muss jedoch kleiner sein als die halbe Wellenlänge. Dieser Zusammenhang ist ein Analogon zum zeitlichen Abtasttheorem und wird deshalb als räumliches Abtasttheorem bezeichnet,

$$f_{\max} < \frac{c}{2\, d_{il}}. \tag{2.14}$$

Wobei d_{il} den Abstand zwischen zwei Sensoren i und l definiert. Wenn Gleichung (2.14) nicht eingehalten wird, können räumliche Aliasingartefakte (engl. spatial aliasing) auftreten.

Geht man nun von einer Abtastfrequenz von z.B. $f_s = 8\,\text{kHz}$ aus, dürfte der Abstand zwischen zwei Sensoren i und l nicht größer als $d_{il} = 4,2\,\text{cm}$ sein. Dies widerstrebt aber oft den technischen Randbedingungen in Anwendungsgebieten der TDoA, denn dort ist ein möglichst großer Sensorabstand, zur Verbesserung der Richtungsvorhersage, gewünscht (vgl. Abschnitt 2.3.3). Da in dieser Arbeit als Quellen aber nur Sprecher von Interesse sind und ebenso eine unverfälschte Übertragung des aufgenommenen Signals nicht von Interesse ist, kann dieses Problem durch einige Vereinfachungen umgangen bzw. abgemildert werden. Auf Grund der Eigenschaften von Sprache (vgl. Abschnitt 2.2) liegt das Hauptaugenmerk bei der TDoA bei Frequenzen von $50\,\text{Hz} \leq f_0 \leq 400\,\text{Hz}$. Ausgehend von diesen Frequenzen kann der Sensorabstand wesentlich größer bemessen werden (Gleichung (2.14)) und die für die TDoA notwendige Laufzeitdifferenz erhöht sich ebenso [Roe07].

[3] Die Schallgeschwindigkeit für Luft wird in dieser Arbeit mit $c = 343\,\text{m/s}$ definiert.

In Tabelle 2.1 sind für einige Grundfrequenzen exemplarisch die zugehörigen, maximalen Mikrofonabstände und die daraus folgenden größtmöglichen Laufzeitunterschiede,

$$\tau_{\max} = \frac{d_{il}}{c}, \tag{2.15}$$

zusammengefasst.

Tabelle 2.1: Zusammenhang von Grundfrequenz f_0 und möglichen Sensorabstand d_{il} nach Gleichung (2.15), sowie den daraus resultierenden größten Laufzeitunterschied zwischen den Sensoren nach Gleichung (2.15). Der Sensorabstand ist nur durch in diesem Abschnitt beschriebenen, besonderen Annahmen möglich.

f_0 /Hz	d_{il} /m	$\tau_{\max}$ /ms
50	$\leq 3,43$	10
200	$\leq 0,86$	2,5
400	$\leq 0,43$	1,25

Die Festlegung des maximalen Sensorabstandes anhand der Grundfrequenz ermöglicht einen praktikablen Kompromiss zwischen der Genauigkeit der Winkelauflösung der TDoA und der Verschlechterung durch räumliches Aliasing [Vol10].

2.3.3 Zusammenhang von Laufzeitdifferenz und Einfallsrichtung

Um mittels des gemessenen Laufzeitunterschiedes zwischen zwei Sensoren i und l auf den Einfallswinkel zu schließen, müssen die geometrischen Betrachtungen aus Abbildung 2.13 verwendet werden. Über eine trigonometrische Funktion

$$\hat{\varphi}_0 = \arccos\left(\frac{\hat{\tau}_{il} \cdot c}{d_{il}}\right) \tag{2.16}$$

kann der geschätzte Einfallswinkel[4] berechnet werden. Dieser Zusammenhang von Laufzeitunterschied und Einfallsrichtung ist nur dann gültig, wenn das Schallfeld als ebene Welle betrachtet werden kann. Eine ebene Welle ist gegeben, wenn man sich im Fernfeld der Signalquelle befindet und der Abstand d_{il} der Mikrofone m_i, m_l wesentlich kleiner ist, als der Abstand der Signalquelle zu den Mikrofonen [Kut07]. Für eine untere Frequenz von $f = 50\,\text{Hz}$ kann ab einem Abstand von $d = 1,5\,\text{m}$ zwischen den Mikrofonen und Signalquellen von Fernfeldbedingungen ausgegangen werden ($2\pi d/\lambda \gg 1$).

Die Winkelgenauigkeit der Schätzung ist abhängig vom Einfallswinkel φ_0. Über die cos-Funktion besteht ein nichtlinearer Zusammenhang zwischen Einfallswinkel φ_0 und Laufzeitunterschied τ_{il}. Je näher der Einfallswinkel in Richtung der

[4] Die Notation $\hat{}$ kennzeichnet die Berechnung einer Schätzgröße.

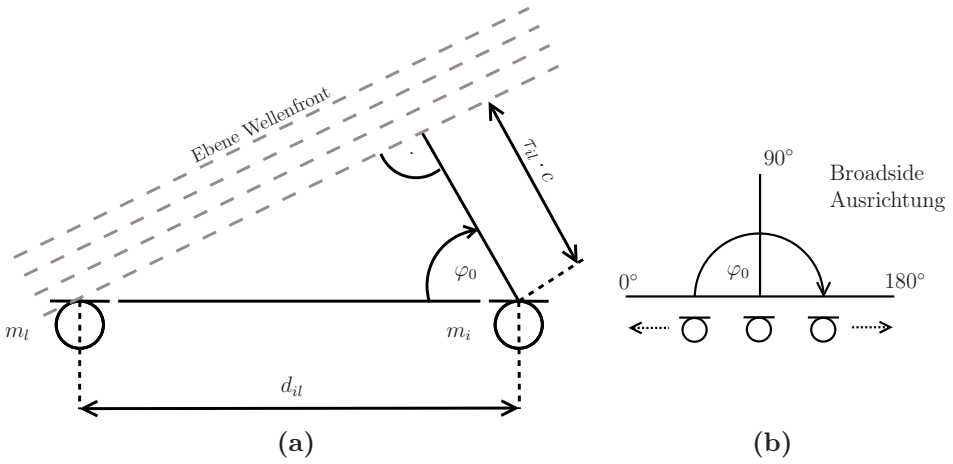


Abbildung 2.13: **(a)** Darstellung des geometrischen Zusammenhangs zwischen dem Einfallswinkel φ_0 und dem Laufzeitunterschied τ unter Annahme einer ebenen Welle. **(b)** Winkeldefinition für die verwendete broadside Ausrichtung des Mikrofon-Arrays.

Mikrofonebene zeigt $\varphi_0 \to 0° \vee 180°$, desto größer ist die Winkelungenauigkeit. Bei Einfallswinkeln nahe $\varphi = 90°$ bildet ein Abtastwert einen kleineren Zeitunterschied ab als bei einem Einfallswinkel nahe der Mikrofonebene. Der durch Gleichung (2.16) bestimmte Winkel kann zudem uneindeutig sein. Sofern nicht durch A priori Wissen eine Einschränkung gemacht werden kann, besitzt die Winkelangabe φ_0 zwei Bedeutungen,

$$\varphi_{0,\text{real}} = \varphi_0 \quad \vee \quad -\varphi_0. \tag{2.17}$$

Diese Vorne-Hinten Verwechselung kann unterdrückt werden, indem z.B. die Mikrofone vor einer Wand plaziert werden oder aber mehr als zwei Mikrofone , die nicht in einer Linie angeordnet sind, ausgenutzt werden. Der Abstand der Signalquelle zu den Mikrofonen kann allein mit einem Mikrofonpaar nicht geschätzt werden. Die Signalquelle kann sich auf jedem beliebigen Punkt der, vom Winkel φ_0, aufgestellten Geraden befinden. Für eine genauere Schätzung der Quellenposition sind weitere Mikrofone nötig, um aus der Kombination mehrerer Mikrofonpaare eine Schätzung der Quellenposition im Raum vorzunehmen.

2.3.4 Zeropadding

Die zeitdiskrete Signalverarbeitung der Mikrofonsignale $x_i[k]$ erfolgt in Blöcken der Länge N, erst somit ist eine diskrete Echtzeitverarbeitung des Signals möglich. Bei der Sprachsignalverarbeitung kann zusätzlich, auf Grund der zeitlichen Änderung von Sprache (vgl. Abschnitt 2.2) bei geeigneter Wahl der Blocklänge N[5], das Nutzsignal in einem Block als stationär angesehen werden. Die Blocklänge sollte für eine praktische Anwendung dabei einer Zweierpotenz entsprechen. So kann bei der Transformation vom Zeitbereich in den Frequenzgang und umgekehrt die

[5] typischerweise 20-30 msec

schnellere Fast Fourier Transformation (FFT) anstatt der DFT verwendet werden[6]. Die Genauigkeit der TDoA Schätzung hängt auch von der zeitlichen Auflösung des digitalen Signals ab. Diese ist zum einen durch die Abtastrate f_s, aber bei einer Schätzung über den Frequenzbereich auch durch die Blöcklänge N bestimmt. Eine Vergrößerung der Blocklänge N zur Vebesserung der Auflösung des Signals hat eine größere Verzögerung der Echtzeitverarbeitung und dem Verlust der Stationaritätsannahme des Nutzsignals zur Folge. Eine Möglichkeit um die Auflösung im Zeitbereich oder Frequenzbereich zu Verbessern besteht im Zeropadding. In diesem Abschnitt werden die Folgen des Zeropadding im Zeitbereich und Frequenzbereich erläutert.

In Abbildung 2.14 ist die Auswirkung des Zeropadding im Zeitbereich dargestellt. Ausgehend von einem Rechtecksignal

$$x[k] = \begin{cases} 1 & \text{, für } 0 \leq k \leq N-1 \\ 0 & \text{, sonst} \end{cases}, \tag{2.18}$$

der Länge $N = 16$, zeigt Abbildung 2.14b die DFT,

$$\text{DFT}: \quad \mathsf{x}[n] = \sum_{k=0}^{N'-1} x[k] \exp^{-j2\pi nk/N'}, \qquad n = 0, 1, 2, \ldots, N'-1, \tag{2.19}$$

mit und ohne Zeropadding. Die neue Blocklänge N' ist dabei definiert als

$$N' = \upsilon \cdot N. \tag{2.20}$$

Bei der Transformation des Signals $x[k]$ nach Gleichung (2.19) erhält man demnach eine spektrale Repräsentation der Länge N'. Wobei für ein reelles Zeitsignal nur Abtastwerte von $n = 0$ bis $n = N'/2+1$ eine Information enthalten. Die zweite Hälfte des Spektrums entspricht einer konjugiert komplexen Version der ersten Hälte und kann jederzeit aus den Elementen $n = 0, \ldots, N'/2 + 1$ rekonstruiert werden.

Der Faktor υ bestimmt die Anzahl der an das Zeitsignal angefügten Nullen. Für $\upsilon = 1$ werden keine Nullen an das Ausgangssignal angehängt. Mit den zusätzlichen Nullen werden keine neuen Informationen beigefügt. Bereits das Spektrum $\mathsf{x}[n]$ mit der Blocklänge N enthält alle Informationen von $x[n]$ und würde zur vollständigen Rekonstruktion ausreichen. Das Anfügen von Nullen führt lediglich zu einer interpolierten Darstellung [PM96, KJ05]. Im Falle eines Rechtecksignals

[6] Es existieren auch FFT Algorithmen für Längen abweichend von Zweierpotenzen, meist werden jedoch Algorithmen verwendet, die für Längen aus Zweierpotenzen optimiert sind.

für $x[k]$ ergibt die DFT eine Sinc-Funktion,

$$\mathsf{x}[n] = \sum_{k=0}^{N'-1} x[k] \exp^{-j2\pi nk/N'} \tag{2.21}$$

$$= \sum_{k=0}^{N-1} x[k] \exp^{-j2\pi nk/N'} \tag{2.22}$$

$$= \sum_{k=0}^{N-1} \exp^{-j2\pi nk/N'} \text{ (geometrische Reihe)} \tag{2.23}$$

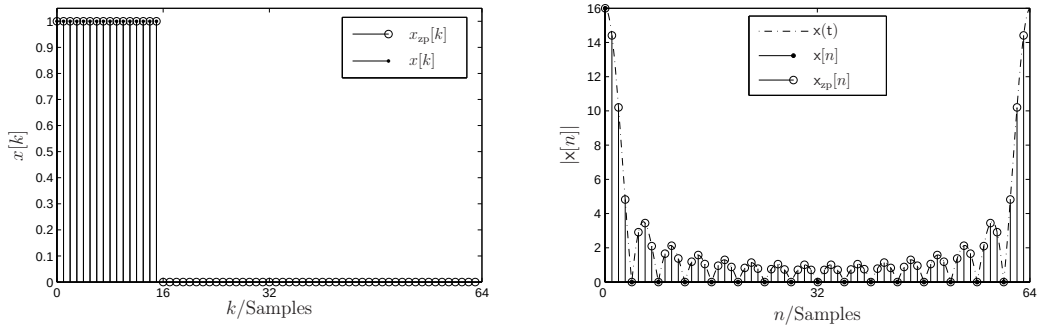
$$= \frac{1 - \exp^{-j2\pi nN/N'}}{1 - \exp^{-j2\pi n/N'}}, \qquad n = 0, 1, 2, \ldots, N' - 1 \tag{2.24}$$

$$\mathsf{x}[n] = \frac{\sin(\pi nN/N')}{\sin(\pi n/N')} \exp^{-j\pi n(N'-1)/N}. \tag{2.25}$$

Für den Fall $\upsilon = 1$ also $N' = N$ enthält die DFT eines Rechtecksignals bis auf $\mathsf{x}[0] = N$ nur Nullen. Die spektrale Repräsentation (vgl. Abbildung 2.14b) stellt somit ein nicht sehr aussagefähiges Bild dar. Mit einer Verbesserung der Auflösung durch Zeropadding ist das dargestellte Bild wesentlich aussagekräftiger. Das Zeropadding erzeugt eine detailliertere Darstellung durch Interpolation von Zwischenwerten. Die diskrete, normierte Kreisfrequenz,

$$\Omega[n] = \frac{2\pi n}{N' f_s} = \frac{2\pi n}{\upsilon N f_s}, \qquad n = 0, 1, 2, \ldots, N' - 1, \tag{2.26}$$

wird um den Faktor υ genauer dargestellt.



(a) Abtastpunkte eines Rechtecksignals im Zeitbereich.

(b) DFT des Rechtecksignals mit sowie ohne Zeropadding.

Abbildung 2.14: Zeropadding im Zeitbereich. **(a)** Abgetastetes Rechtecksignal $x[k]$ und $x_{\text{zp}}[k]$ (mit und ohne Zeropadding). Faktor υ beträgt 4, die Blocklänge ist auf $N = 16$ gesetzt. **(b)** Betrag des Ergebnisses der DFT mit ($\mathsf{x}_{\text{zp}}[k]$) sowie ohne ($\mathsf{x}[k]$) Zeropadding. Die gestrichelte Line stellt den kontinuierlichen Verlauf dar.

Das Prinzip des Zeropadding kann auch analog im Frequenzbereich angewendet werden. Dabei werden die Nullen jedoch nicht am Ende eines Blocks angefügt,

sondern in dessen Mitte. Dies geschieht auf Grund der Eigenschaft der DFT, die negativen Frequenzindizes zyklisch verschoben an den Stützstellen $N/2+2 \ldots N-1$ darzustellen, also am Ende des positiven, diskreten Frequenzbereich. Denn wie schon beschrieben enthalten nur $N/2+1$ Samples der Frequenzdarstellung Informationen. Somit müssen der ersten Hälfte von $\mathsf{x}[n]$ $(\upsilon-1)N/2$ Nullen angehängt werden, um eine vollständige IDFT des Signals zu ermöglichen. Die Abbildung 2.15 stellt das Prinzip des Zeropadding im Frequenzbereich durch eine IDFT,

$$\text{IDFT}: \quad x[k] = \frac{1}{N} \sum_{k=0}^{N'-1} \mathsf{x}[n] \exp^{j2\pi nk/N'}, \qquad k = 0, 1, 2, \ldots, N'-1, \tag{2.27}$$

eines um n_0 verschobenen Deltapulses,

$$\mathsf{x}[n] = \begin{cases} \frac{N}{2} & \text{, für } n = \frac{N'}{2} \pm \left(\frac{N'}{2} - n_0\right) \\ 0 & \text{, sonst} \end{cases}, \tag{2.28}$$

dar. Die sich dabei über n_0 einstellende Kreisfrequenz ergibt sich aus

$$\Omega_{n_0} = 2\pi \frac{n_0}{N}, \tag{2.29}$$

was im Zeitbereich einer periodischen Schwingung mit Ω_{n_0} entspricht.

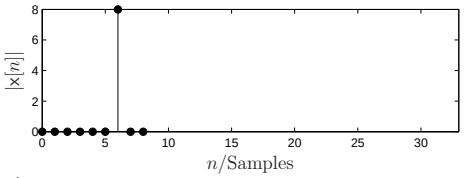
Abbildung 2.15c stellt die IDFT von Gleichung (2.28) dar:

$$x[k] = \frac{1}{N}\frac{N}{2}\left(\exp^{j\left(\frac{N}{2}-\left(\frac{N'}{2}-n_0\right)\right)2\pi k/N'} + \exp^{j\left(\frac{N}{2}+\left(\frac{N'}{2}-n_0\right)\right)2\pi k/N'}\right) \tag{2.30}$$

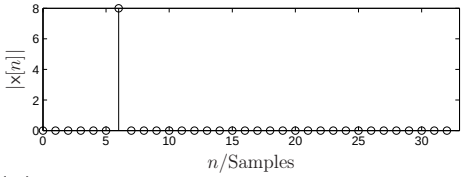
$$= \frac{1}{2} \underbrace{\exp^{j\pi k}}_{(-1)^k} \underbrace{\left(\exp^{-j\left(\frac{N'}{2}-n_0\right)2\pi k/N'} + \exp^{j\left(\frac{N'}{2}-n_0\right)2\pi k/N'}\right)}_{2\cos\left(\left(\frac{N'}{2}-n_0\right)2\pi k/N'\right)} \tag{2.31}$$

$$x[k] = (-1)^k \cos\left(\pi k \left(1 - \frac{2\pi n_0}{N'}\right)\right), \qquad k = 0, 1, 2, \ldots, N'-1. \tag{2.32}$$

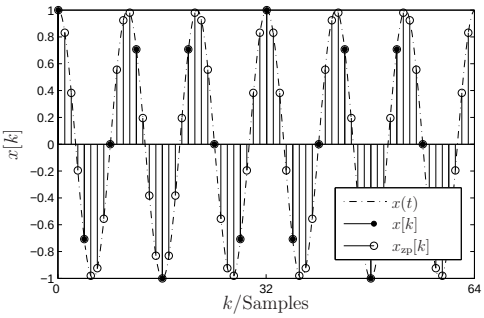
Das Zeropadding ist eine sinnvolle Technik, wenn es z.B. darum geht, Maxima in einem transformierten Signal zu suchen, wie es auch bei der TDoA der Fall ist (vgl. Abschnitt 2.3.5). Durch diese Interpolation kann der Index der Maxima genauer bestimmt werden, wenn sie ohne Zeropadding z.B. zwischen zwei Abtastwerten liegen würden.



(a) positives Spektrum ohne Zeropadding.



(b) positives Spektrum mit Zeropadding.



(c) IDFT des Deltapulses.

Abbildung 2.15: Zeropadding im Frequenzbereich. Einseitiges Spektrum für positive Frequenzen, **(a)** ohne Zeropadding ($\mathsf{x}[n]$, **(b)** mit Zeropadding ($\mathsf{x}_{\mathrm{zp}}[n]$). Der Faktor υ beträgt 4, die Blocklänge ist $N = 16$. Die Nullen sind in die Mitte des Spektrums eingefügt. **(c)** IDFT des konjugiert komplexen Deltapulses mit $x_{\mathrm{zp}}[k]$ sowie ohne $x(t)$. Die gestrichelte Linie stellt den kontinuierlichen Verlauf dar.

2.3.5 Kreuzkorrelation

Die Kreuzkorrelationsfunktion (KKF) ist eine effektive Möglichkeit, den relativen Laufzeitunterschied zweier Signale zu schätzen. Dabei werden die Signale mit zwei räumlich getrennten Mikrofonen aufgenommen. Eine Laufzeitschätzung ist auf Grund der Tatsache möglich, dass die Signale bei dem hier betrachteten Anwendungsgebiet ursprünglich aus einer Signalquelle stammen. Der Zeitunterschied ergibt sich aus den unterschiedlich langen Laufwegen des Signals zu den Sensoren. Ausgehend von einem Modell der freien Schallausbreitung (Abschnitt 2.3.1), können die aufgenommenen Mikrofonsignale (entsprechend Gleichung (2.9)) folgenderweise beschrieben werden:

$$x_i[k] = a_i s[k - \kappa_i] + q_i[k] \tag{2.33}$$

$$x_l[k] = a_l s[k - \kappa_l] + q_l[k]. \tag{2.34}$$

Dabei stellen die Signale an den Mikrofonen $x[k]$ eine um κ verzögerte und a gedämpfte Version des Originalsignals dar, welches von einem, zum Nutzsignal unkorrelierten, Rauschen $q[k]$ überlagert ist. Aus dem Mikrofonsignal kann jedoch keine Aussage über die Dämpfungsfaktor a gewonnen werden.

Grundlegendes zur Korrelation findet sich im Abschnitt 2.2.3 zur Grundfrequenzerkennung. Setzt man die Mikrofonsignale aus Gleichung (2.33) in die Gleichung (2.4) zur Bestimmung des Erwartungswertes der Korrelation ein, so ergibt sich unter der Annahme das Nutzsignalanteil und Rauschanteil unkorreliert sind folgende Schreibweise:

$$r_{x_i x_l}[\kappa] = \mathrm{E}\left\{\left(a_i s[k - \kappa_i] + q_i[\kappa]\right)\left(a_l s[k + \kappa - \kappa_l] + q_l[k + \kappa]\right)\right\} \tag{2.35}$$

$$\approx \mathrm{E}\left\{a_i a_l s[k - \kappa_i] s[k + \kappa - \kappa_l] + q_i[\kappa] q_l[\kappa + \kappa_l]\right\}. \tag{2.36}$$

Bei der Ausmultiplikation von Gleichung (2.35) entstehen zusätzliche Kreuzkorrelationstherme. Diese Therme wurden hier beim Übergang von Gleichung (2.35) zu Gleichung (2.36) nicht extra angegeben, denn sie können zu Null gesetzt werden. Begründet ist dies durch die, bei dem Modell der freien Schallausbreitung, festgelegte Unkorreliertheit der Nutzsignale mit dem Rauschen. Die sich aus Gleichung (2.36) ergebende Schätzung lässt sich wie folgt ermitteln:

$$\hat{r}_{x_i x_l}[\kappa] = \frac{1}{N}\sum_{k=0}^{N-1} a_i a_l s[k-\kappa_i]s[k+\kappa-\kappa_l] + \underbrace{\frac{1}{N}\sum_{k=0}^{N-1} q_i[\kappa]q_l[\kappa+\kappa_l]}_{\hat{r}_{q_i q_l}[\kappa]}. \tag{2.37}$$

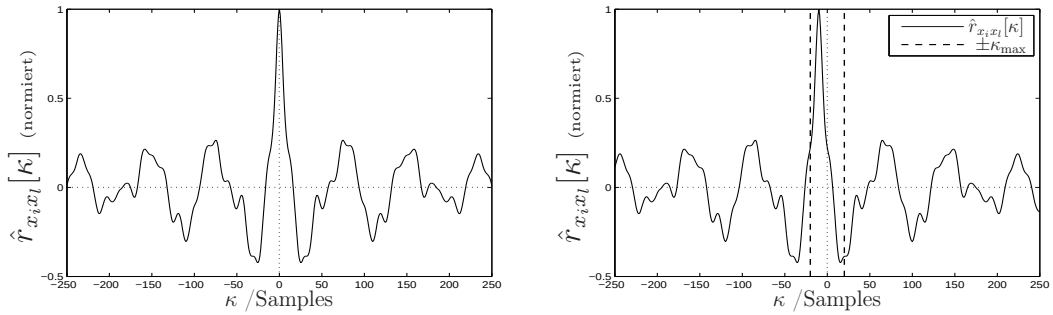
Die hintere Summe aus Gleichung (2.37) kann als Kreuzkorrelation $\hat{r}_{q_i q_l}[\kappa]$ des additiven Rauschens q_i und q_l beschrieben werden. Im folgenden Schritt werden die einzelnen Laufzeitverzögerungen κ_i und κ_l durch den relativen Laufzeitunterschied κ_{il} zwischen den Mikrofonen mit

$$\kappa_{il} = \kappa_l - \kappa_i \tag{2.38}$$

substituiert:

$$\hat{r}_{x_i x_l}[\kappa] = \frac{1}{N} a_i a_l \sum_{k=0}^{N-1} s[k]s[k+\kappa-\kappa_{il}] + \hat{r}_{q_i q_l}[\kappa] \tag{2.39}$$

$$= a_i a_l \cdot \hat{r}_{ss}[\kappa-\kappa_{il}] + \hat{r}_{q_i q_l}[\kappa]. \tag{2.40}$$



(a) Autokorrelation eines Sprachsignalblocks. **(b)** Kreuzkorrelation eines Sprachsignalblocks mit $f_s = 24\,\text{kHz}$ und $d_{il} = 0,28\,\text{m}$.

Abbildung 2.16: **(a)**Autokorrelationsfunktion eines Sprachsignalblocks. **(b)** Kreuzkorrelationsfunktion des gleichen Sprachsignalblocks wie in **(a)**, jedoch wurde ein Kanal um 10 Samples verzögert, dies entspricht nach Gleichung (2.16) einem Winkel von $\varphi = 59°$ (bei $d_{il} = 0,28\,\text{cm}$). Die maximale Verzögerung ist hier nach Gleichung (2.42) mit 20 Sampels angegeben.

Die Kreuzkorrelation der Signale x_i und x_l entspricht somit einer verschobenen Autokorrelation von einem Nutzsignal $s[k]$ plus einer Kreuzkorrelation des additiven Rauschens. Das Maximum der Korrelation ist somit ebenfalls um die relative

Laufzeitverschiebung κ_{il} vom Ursprung der Korrelationsfolge verschoben. Aus der Ermittlung der Verschiebung des Maximums der Korrelationsfolge um κ_{il} relativ zu $\kappa = 0$ lässt sich demzufolge die Laufzeitdifferenz des Signals zwischen zwei Mikrofonen schätzen,

$$\hat{\tau}_{il} = \kappa_{il}/f_s. \tag{2.41}$$

Der Schalleinfallswinkel lässt sich anschließend mit Gleichung (2.16) bestimmen.

Um die Maximumsuche effektiv zu gestalten, kann diese auf einen eingeschränkten Bereich innerhalb der Korrelation bezogen werden. Der maximal mögliche Bereich innerhalb dessen sich das Maximum befinden muss, ist durch den Mikrofonabstand festgelegt:

$$\kappa_{\max} = \pm\frac{d_{il}\, f_s}{c}. \tag{2.42}$$

In Abbildung 2.16b ist dieser Bereich durch gestrichelte Linien dargestellt. Maxima, die sich außerhalb dieses Suchbereiches befinden, können nicht durch den Laufzeitunterschied von Mikrofon i zu l stammen.

Effektive Berechnung der Korrelation im Frequenzbereich

Auf Grund der Effektivität der Fast Fourier Transformation (FFT) kann die Berechnung der KKF auch mit Hilfe der Bestimmung des Leistungsdichtespektrums (LDSs) vorgenommen werden. Die LDS ist die fouriertransformierte Korrelation im Frequenzbereich,

$$\Phi[n] = \sum_{\kappa=0}^{N-1} r[\kappa] e^{-j\kappa\frac{2\pi}{N}n} \qquad , n = 0 \ldots N-1. \tag{2.43}$$

Der Vorteil dieser Methode ist es, dass sich die Faltung der Signale im Zeitbereich zu einer Multiplikation im Frequenzbereich reduziert und somit eine schnellere Berechnung ermöglicht. Entsprechend der Unterscheidung von KKF und AKF werden die transformierten Signale als Autoleistungsdichtespektrum (ALDS) bzw. Kreuzleistungsdichtespektrum (KLDS) bezeichnet. Analog zur Bestimmung von $r_{x_i x_l}[\kappa]$ aus den Zeitsignalen $x_i[k]$ und $x_l[k]$, errechnet sich das LDS aus dem Erwartungswert der Signalspektren,

$$\Phi_{\mathsf{x}_i\mathsf{x}_l}[n] = \lim_{N\to\infty} \frac{1}{N} \mathrm{E}\{\mathsf{x}_i[n]\mathsf{x}_l{}^*[n]\}. \tag{2.44}$$

Die Schätzung der Leistungsdichtespektren muss bei stochastischen Prozessen theoretisch über einen unendlichen langen Signalvektor erfolgen. Bei der Sprachsignalverarbeitung ist jedoch auf Grund der Kurzzeitstationarität von Sprache eine rekursive Mittelung der Leistungsdichtespektren allgemein üblich [Bitzer01]:

$$\hat{\Phi}_{\mathsf{x}_i\mathsf{x}_l}[n,\ell] = \alpha\hat{\Phi}_{\mathsf{x}_i\mathsf{x}_l}[n,\ell-1] + (1-\alpha)\mathsf{x}_i[n,\ell]\mathsf{x}_l{}^*[n,\ell]. \tag{2.45}$$

Die Laufzahl ℓ in Gleichung (2.45) gibt die aktuell verarbeitete Blocknummer an. Der Glättungsfaktor α ergibt sich aus dem Blockvorschub V, dem Glättungszeitraum t_s und der Abtastfrequenz f_s[7],

$$\alpha = e^{-\frac{V}{t_s f_s}}. \tag{2.46}$$

Um über die LDS auf die Korrelation zu gelangen, muss diese noch durch eine IDFT in den Zeitbereich transformiert werden,

$$\hat{r}[k] = \frac{1}{N}\sum_{n=0}^{N-1} \hat{\Phi}[n] e^{jk\frac{2\pi}{N}n} \qquad , k = 0 \ldots N-1\,. \tag{2.47}$$

Durch die Berechnung der Korrelation über die LDS kann das Zeropadding im Frequenzbereich zur Interpolation angewendet werden. Die Blocklänge aus Gleichung (2.47) muss dann, entsprechend Abschnitt 2.3.4, zu N' angepasst werden.

Zum Vergleich der beiden dargestellten Verfahren zur Ermittlung der KKF wird im Folgenden grob deren numerischer Rechenaufwand bestimmt. Dabei werden Additionen/Subtraktionen nicht berücksichtiget, da angenommen werden kann, dass deren Aufwand vernachlässigbar klein ist [CBH06]. Es wird festgelegt, dass die Multiplikation zwei komplexer Zahlen mit vier reellwertigen Multiplikationen ausgedrückt werden kann und die Multiplikation von reellwertigen und komplexwertigen Zahlen zwei reellwertigen Multiplikationen entspricht [CBH06]. Der numerische Rechenaufwand einer Faltung, zweier Signalblöcke der Länge N, im Zeitbereich ergibt auf Grundlage von Gleichung (2.5) dann einen Wert von

$$\mathcal{R}_{\text{zeit}} = N^2 + N \tag{2.48}$$

reellwertigen Multiplikationen für jedes zu berechnende Sample. Um den groben Rechenaufwand für die Ermittlung der Korrelation über das LDS anzugeben, muss der Rechenaufwand für die schnelle Faltung (FFT und IFFT) bekannt sein, welcher nach [KK09] für einen Radix-2 Algorithmus mit $\mathcal{R} = \frac{N}{2}\log_2(\frac{N}{2})$ angegeben ist. Der Aufwand für Gleichung (2.45) kann mit den beschriebenen Festlegungen grob mit $8N$ bzw. $4N$ (für konjugiert komplexe Signale) bestimmt werden. Für die Ermittlung der KKF aus dem LDS besteht ein Rechenaufwand pro Signalblock von drei Transformationen und einer LDS Berechnung nach Gleichung (2.45):

$$\mathcal{R}_{\text{freq}} = 3\frac{N}{2}\log_2\left(\frac{N}{2}\right) + 8N \tag{2.49}$$

Durch die Gleichungen (2.48) und (2.49) ist zu erkennen, dass der Aufwand für die Faltung quadratisch mit der Blocklänge anwächst, wohingegen die Berechnung im Frequenzbereich nur einen logarithmischen Zusammenhang aufweist. Somit gilt die Berechnung über das LDS trotz der notwendigen Transformationen für praktikable Blocklängen N als effektiver im Vergleich zur Faltung im Zeitbereich.

[7] In dieser Arbeit wurde ein Vorschub von $V = 50\%$ und eine Glättungszeit von $t_s = 0,1\,\text{s}$ verwendet.

2.3.6 Generalized Cross-Correlation

Die Generalized Cross-Correlation (GCC) ist nach [KC76] eine KKF, bei der die Eingangssignale $x_i[k]$ und $x_l[k]$ vor der eigentlichen Korrelationberechnung noch mit den Filtern $z_i[k]$ und $z_l[k]$ vorverarbeitet werden. Ziel der Vorverarbeitung ist es, die Ausprägung des Maximums in der Korrelationsfolge zu verstärken. Im Idealfall der freien Schallausbreitung ohne Rauschen (siehe Abschnitt 2.3.1) würde der Einfallswinkel durch einem verschobenen Deltapuls im Zeitbereich repräsentiert werden. In Abbildung 2.17 ist eine GCC eines Sprachsignals ohne zusätzliches Rauschen und Reflexionen dargestellt.

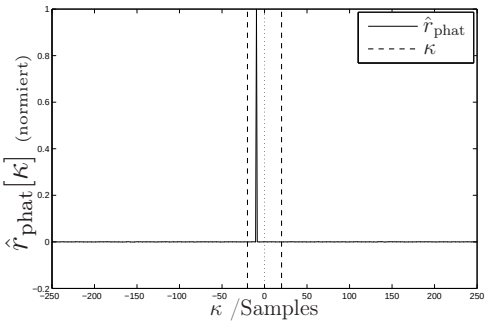


Abbildung 2.17: GCC Funktion mit Phat Transformation eines Sprachsignals ohne Rauschen und Refelexionen. Das Signal entspricht dem aus Abbildung 2.16, $f_s = 24\,\text{kHz}$ und $d_{il} = 0,28\,\text{m}$.

Eine oft angewendete Filterfunktion stellt dabei die Phat dar. Dies ist ein sogenanntes Pre-Whitening Filter, welches die Amplitude des Leistungsspektrums auf konstant eins umformt. Die Transformation eines Deltapulses in den Frequenzbereich ist ein weißes Sprektrum mit linear abfallender Phase, über die die winkelbestimmende Laufzeitverzögerung gekennzeichnet ist. Abbildung 2.18 stellt die DFT zweier verschobener Deltapulse beispielhaft dar.

Um das Mikrofonsignal $x[k]$ mit der Phat Transformation zu bearbeiten, muss dies mit dessen eigenem inversen Betragsspektrum

$$\mathsf{z}[n] = \frac{1}{|\mathsf{x}[n]|} \tag{2.50}$$

multipliziert werden. Nach Gleichung (2.45)[8] berechnet sich das transformierte LDS mit Phat-Transformation zu:

$$\hat{\phi}^{\text{GCC}}_{x_i,x_l}[n] = \mathsf{z}_i[n]\,\mathsf{x}_i[n]\;\mathsf{z}_l{}^*[n]\,\mathsf{x}_l{}^*[n]. \tag{2.51}$$

In Gleichung (2.51) können die Filter $\mathsf{z}[n]$ durch Gleichung (2.50) ersetzt werden und die Konjugation von $\mathsf{z}_l[n]$ kann vernachlässigt werden, da $\mathsf{z}_l[n]$ eine rein

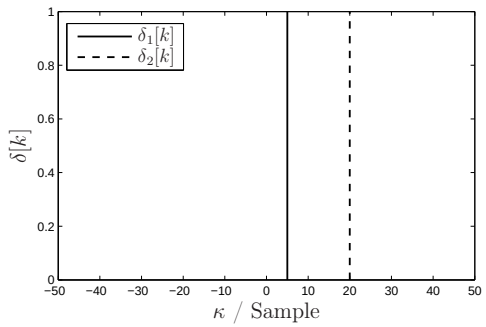
[8] Die rekursive Glättung aus Gleichung (2.50) wird hier aus Gründen der Lesbarkeit vernachlässigt, ist aber für eine praktische Anwendung der GCC ebenso zu berücksichtigen.

reellwertige Funktion ist.

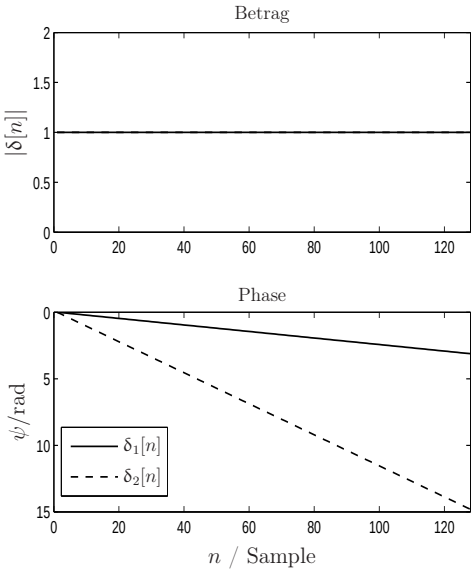
$$\hat{\phi}^{\mathrm{GCC}}_{x_i,x_l}[n] = \frac{\mathsf{x}_i[n]\,\mathsf{x}_l{}^*[n]}{|\mathsf{x}_i[n]\,\mathsf{x}_l[n]|} \tag{2.52}$$

$$\hat{r}_{x_i x_l,\mathrm{phat}} = IDFT\left\{\hat{\phi}^{\mathrm{GCC}}_{x_i,x_l}[n]\right\}. \tag{2.53}$$

Durch die Phat-Transformation (weißes Betragsspektrum) geht jedoch die harmonische Struktur der Eingangssignale verloren und es kann keine Grundfrequenzerkennung anhand der GCC vorgenommen werden. Weiterhin werden durch die Vorverarbeitung allen Frequenzen der gleiche Informationsgehalt beigemessen, wodurch das GCC-Verfahren anfällig gegenüber zusätzlichen Rauschanteilen wird (vgl. Abschnitt 5.9).



(a) Zeitdarstellung zweier verschobener Deltapulse.



(b) Spektrum zweier verschobner Deltapulse, aufgetragen in Betrag und Phase.

Abbildung 2.18: **(a)** Zeitdarstellung zweier verschobener Deltapulse mit einem Versatz von 5 bzw. 20 Samples. **(b)** Darstellung der DFT der verschobenen Deltapulse aus **(a)** in Betrag und Phase (in einem Ausschnitt von 125 Samples). Je größer die Verschiebung des Deltapulses im Zeitbereich ist, desto schneller ist die lineare Phasenverschiebung im Frequenzbereich. Der Betrag hingegen bleibt konstant bei eins für alle Frequenzstützstellen. Die Blocklänge der DFT beträgt $N = 1024$.

Kapitel 3

Kombinierte Grundfrequenz- und Richtungsschätzung

In den vorangegangenen Grundlagen wurden Methoden beschrieben, wie die Quellenmerkmale Grundfrequenz und Einfallsrichtung getrennt voneinander berechnet werden können. Dieses Kapitel beschreibt Algorithmen, die eine kombinierte Grundfrequenz- und Richtungsschätzung gestatten und damit eine bessere Analyse der akustischen Szene ermöglichen sollen. Eine Besonderheit dieses Verfahrens ist, dass aus einem eindimensionalen akustischen Signal eine zweidimensionale Darstellung der Grundfrequenz- und Richtungsschätzung erzeugt werden kann [KPW07]. Es kann grundsätzlich zwischen zwei Berechnungsarten für die kombinierte Schätzung unterschieden werden. Zum einen basiert die Schätzung auf der Kreuzkorrelation $r[k]$ im Zeitbereich und zum anderen auf dem Kreuzleistungsdichtespektrum $\phi[n]$ mindestens zweier Mikrofonsignale im Frequenzbereich (Abbildung 3.1). Zu den Kernfunktionen werden zusätzliche Erweiterungen erläutert, wobei hier besonders auf die eigene Phasentransformation, die Erweiterung auf mehr als ein Mikrofonpaar über die Multichannel Cross-Correlation (MCCC)-Integration und der nachträglichen GCC-Phat Gewichtung der kombinierten Schätzung hingewiesen sein soll.

Der Abschnitt 3.1 gibt einen generellen Überblick zu den implementierten Verfahren und deren Zusammenspiel. Im Abschnitt 3.2 und 3.3 werden die Kernfunktionen der kombinierten Grundfrequenz- und Richtungsschätzung behandelt. Im nachfolgenden Abschnitt 3.4 wird auf die mehrkanalige Erweiterung des Algorithmus eingegangen. Abschnitt 3.5 beschäftigt sich mit der Cepstrumsgewichtung und in Abschnitt 3.6 wird auf die Integration einer Filterbank zur möglichen Verbesserung der simultanen Mehrsprecherbestimmung eingegangen. Zuletzt werden in Abschnitt 3.7 die aus der Literatur [HOK08] entnommene und die eigene GCC-Phat Erweiterungen erörtert.

3.1 Signalflussdiagramm

Für einen besseren Überblick ist der Signalflussweg des erstellten Gesamtsystems in Abbildung 3.1 dargestellt. Das Flussdiagramm illustriert das implementierte Zusammenspiel der realisierten Verfahren. Es kann eine grobe Unterteilung in

drei Abschnitte (grau hinterlegt) vorgenommen werden, angefangen mit der Vorverarbeitung, gefolgt von den Kernfunktionen und abschließend mit den Nachverarbeitungen.

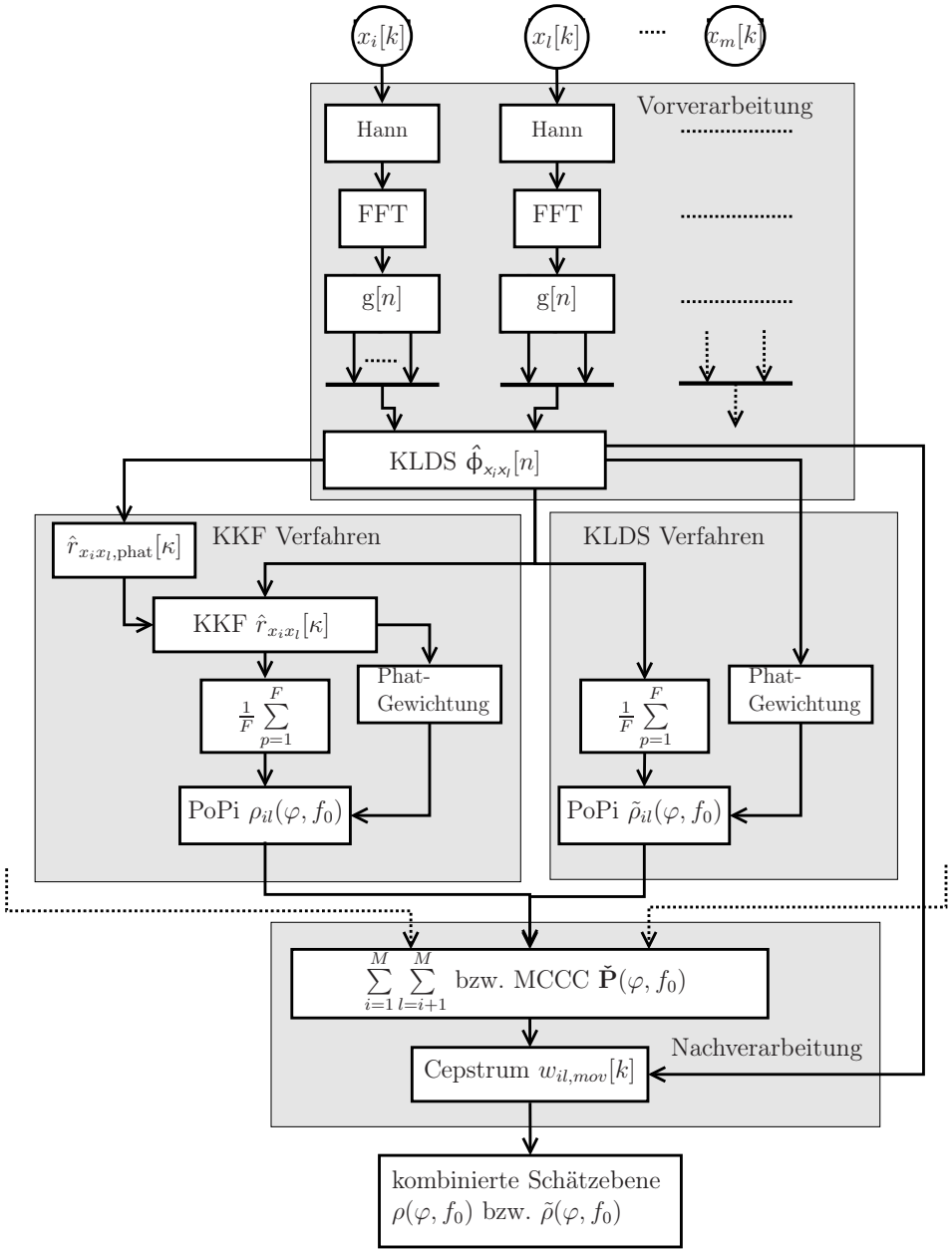


Abbildung 3.1: Zusammenfassendes Signalflussdiagramm zur kombinierten Grundfrequenz- und Richtungsschätzung. Dargestellt sind die beiden Schätzverfahren über die KKF oder das KLDS (grau hinterlegt), sowie alle zusätzlich implementierten Erweiterungen. g[n] symbolisiert eine Unterteilung in F Spektren.

Der Signalweg im Flussdiagramm beginnt mit den Mikrofonsignalen $x_i[k], x_l[k]$. Um den sogenannten Leck-Effekt (engl. leakage) bei der blockweisen Verarbeitung zu reduzieren, werden die Mikrofonsignale in der Vorverarbeitung mit einem Han-

ning-Fenster (auch als Hann-Fenster bekannt [KK09]) multipliziert. Anschließend wird das Zeitsignal durch eine FFT-Filterung in den Frequenzbereich transformiert. Die erste Erweiterung stellt die Unterteilung des Spektrums $\mathsf{x}[n]$ in Frequenzgruppen dar. In diesem Fall wird eine psychoakustisch motivierte Gammatonfiltergruppe verwendet. Nach [KOH08] soll diese eine Verbesserung der Schätzung für Mehrsprechersituationen bewirken (vgl. Abschnitt 3.6 und 5.6). Dabei sind die Filterfunktionen als Bandpass-Gammatonefenster ausgeführt und erzeugen F Spektren, jeweils mit unterschiedlichen Fensterfunktionen multipliziert. Anschließend wird aus den Eingangskanälen das KLDS berechnet. In Abhängigkeit von der Verwendung der Frequenzgruppen erfolgt dies entweder für jede Frequenzgruppe getrennt oder für das komplette Spektrum.

Wie in der Einleitung zu diesem Kapitel schon angedeutet, kann die kombinierte Grundfrequenz- und Richtungsschätzung im Kern auf zwei Arten berechnet werden. Der Signalflussgraph in Abbildung 3.1 trennt sich in einen rechten und linken Pfad auf. Auf der linken Seite ist die Berechnug über die KKF geschildert (vgl. Abschnitt 3.2). Nach der Ermittlung der KKF werden die eventuell vorhandenen Frequenzgruppen zu einer KKF zusammengefasst. Anschließend wird die Grundfrequenz- und Richtungsschätzung für ein Mikrofonpaar ermittelt. Auf der rechten Seite des Signalflussgraphen wird die Schätzung direkt aus dem KLDS ermittelt (vgl. Abschnitt 3.3). Zuvor werden eventuell vorhandene Frequenzgruppenunterteilungen ebenfalls wieder zusammengeführt. Die zusätzlichen Phat-Funktionsgruppen stellen weitere additive Verfahren dar, genauers dazu im Abschnitt 3.7.

Im Anschluss an die Schätzung für die jeweiligen Mikrofonpaare müssen diese durch eine Berechnung der Mittelwerte oder über ein MCCC-Verfahren [CBH03, BCH04] in der Nachverarbeitung zu einer Gesamtschätzung fusioniert werden. Nach [HKO08] kann zusätzlich zu einer Unterdrückung von Mehrdeutigkeiten in der Grundfrequenzschätzung eine auf dem Cepstrum basierende Gewichtung angewendet werden (vgl. Abschnitt 3.5 und 5.6).

Der dargestellte Signalverlauf gibt die maximale Anzahl von verwendbaren und implementierten Verfahren wieder. Dies muss aber nicht zwangsläufig zu den besten Schätzungsresultaten führen. Eine Auswertung der erzielten Resultate findet sich in Abschnitt 5. Der gezeigte Signalverlauf in Abbildung 3.1 gilt für ein Mikrofonpaar, eine Erweiterung auf mehr als zwei Mikrofone ist angedeutet.

3.2 Kombinierte Schätzung über die Kreuzkorrelation

Dieser Abschnitt befasst sich mit der kombinierten Grundfrequenz- und Richtungsschätzung basierend auf der Kreuzkorrelation $\hat{r}_{x_i x_l}[\kappa]$. Auf Grundlage von Gleichung (3.1) wird eine parametrisierte Abtastung der KKF durchgeführt und eine über Grundfrequenz f_0 und Winkel φ aufgespannte Ergebnisebene aufgestellt $\rho_{il}(\varphi, f_0)$ [KPW07].

$$\rho_{il}(\varphi, f_0) = \frac{1}{2P+1} \sum_{p=-P}^{P} \hat{r}_{x_i x_l} \left[\left[p \cdot \underbrace{\frac{2\pi}{\Omega(f_0)}}_{\nu_0(f_0)} + \kappa_{il}(\varphi) \right] \right] \tag{3.1}$$

$$\Omega(f_0) = \frac{2\pi \cdot f_0}{f_s} = \Omega \cdot f_0 \tag{3.2}$$

$$\kappa_{il}(\varphi) = \frac{d_{il} \cdot \cos(\varphi) \cdot f_s}{c}. \tag{3.3}$$

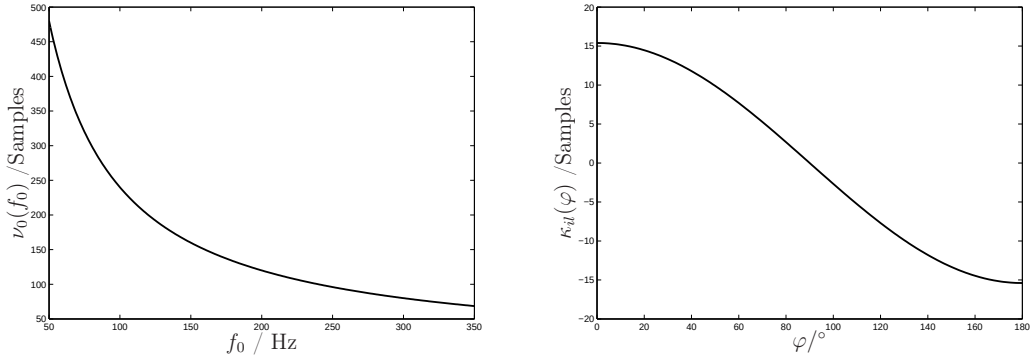
$$\rho'_{il}(\varphi, f_0) = |\rho_{il}(\varphi, f_0)| / \rho_{il,\max}(\varphi, f_0) \tag{3.4}$$

Der Parameter $\nu_0(f_0) = 2\pi/\Omega(f_0)$ definiert den normierten Abstand zwischen zwei Korrelationsmaxima in Samples, hervorgerufen durch die Wellenlänge der Grundfrequenz f_0. Die Anzahl der berücksichtigten Korrelationsmaxima wird mit P festgelegt. Gleichung (3.1) kann dementsprechend als eine Kammabtastung angesehen werden (siehe Abbildung 3.3). Die durch die Einfallsrichtung verursachte Verschiebung der Korrelation vom Nullpunkt $\kappa_{il}(\varphi)$ ist in Gleichung (3.3) gegeben und abhängig vom Mikrofonabstand d_{il}, der Schallgeschwindigkeit c, Abtastfrequenz f_s und dem laufzeitäquivalenten Einfallswinkel φ_0 [KPW07]. Die vorgestellten Betrachtungen gelten streng genommen nur im vereinfachten Fall der Fernfeldannahme, bei kleinerem Mikrofon-Quellen-Abstand muss ein Ausgleich über den Mikrofonabstand d_{il} in Abhängigkeit zur Quellenentfernung erfolgen.

Um eine schnellere Berechnung der Ergebnisebene zu gewähren, können die Parameter für $\nu_0(f_0)$ und $\kappa_{il}(\varphi)$ als Lookup-Tabellen vorbereitet werden. Die verwendete Genauigkeit betrug in dieser Arbeit jeweils 1° für die Winkelauflösung und 1 Hz für die Grundfrequenzauflösung. Die interessierenden Winkel liegen im Bereich von $0° \leq \varphi \leq 180°$, wobei 90° als frontale Einfallsrichtung definiert ist (vgl. Abschnitt 2.3.3). Als Grundfrequenzspektrum kommt der Sprachbereich des Menschen von ca. 50 Hz bis ca. 400 Hz in Betracht (vgl. Abschnitt 2.2.2).

In Abbildung 3.2 sind die Verläufe der Funktionen $\nu_0(f_0)$ und $\kappa_{il}(\varphi)$ aufgetragen. Beim Verlauf der winkelabhängigen Korrelationsverschiebung $\kappa_{il}(\varphi)$ in der rechten Abbildung 3.2b ist die Winkelrepräsentation als Versatz in Abtastwerten dargestellt. Mit größer werdendem Einfallswinkel $\varphi \to 0°$bzw.$\varphi \to 180°$ nimmt die Genauigkeit der Winkelauflösung ab. Bei einem angenommenen gleichen Winkelabstand ist der Laufzeitunterscheid für große Einfallswinkel $\varphi \to 0°$bzw.$\varphi \to 180°$ geringer als für kleinere Einfallswinkel $\varphi \to 90°$ (vgl. Abbildung 2.13). Auch beim Verlauf der Kammbreite (Abbildung 3.2a) ist eine von der Grundfrequenz abhängige Genauigkeit festzustellen.

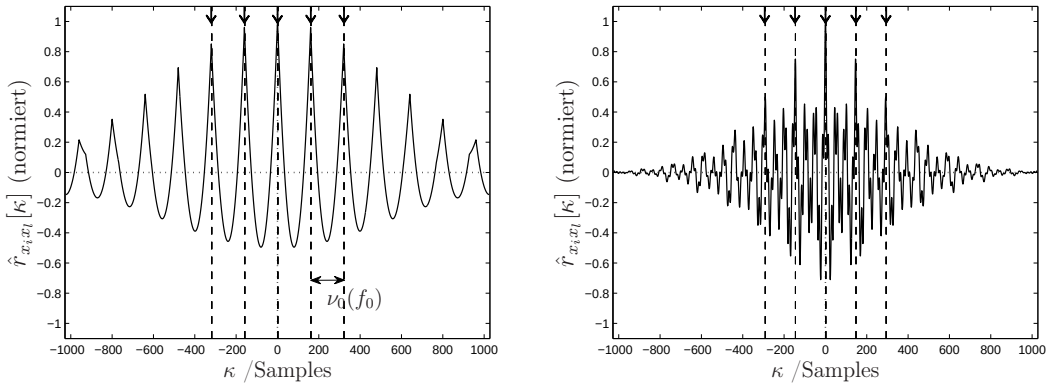
Die Abbildung 3.3 zeigt das Ergebnis der Kammabtastung Kammabtastung der Korrelation $\rho_{il}(\varphi, f_0)$ am Beispiel eines Sägezahnsignals (3.3a) sowie einer Sprachaufnahme (3.3b). Es ist darauf zu achten, dass ein passender Zusammenhang zwischen Mikrofonabstand d_{il}, Blockgröße N, Anzahl der betrachteten Korrelationsmaxima P und interessierenden Grundfrequenzen f_0 besteht. Durch eine



(a) **(b)**

Abbildung 3.2: **(a)** Verlauf der Korrelationsmaximaabstände $\nu_0(f_0)$, **(b)** Versatz $\kappa_{il}(\varphi)$ der Kreuzkorrelation für einen Mikrofonabstand $d = 22\,\text{cm}$ und einer Abtastrate $f_s = 24\,\text{kHz}$.

große Anzahl P Korrelationsmaxima und kleiner Grundfrequenz f_0 können relevante Sampleindizes außerhalb der Blockgröße N liegen. Bei der zur Quelle korrespondierenden Kammbreite und Nulllinienversatz resultiert die Summe der Abtastpunkte in einem Maximum und repräsentiert somit die Schätzung der Signalquelle in der Ergebnisebene.

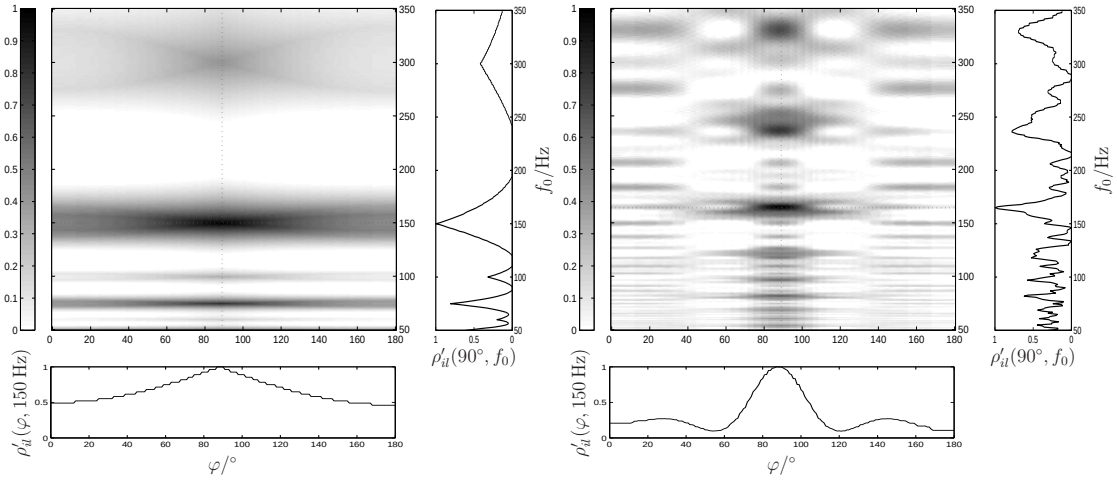


(a) Korrelation eine Sägezahnsignals **(b)** Korrelation eines Sprachsignals

Abbildung 3.3: Autokorrelationen (Annahme: $x_i[k] = x_l[k] \rightarrow \varphi_0 = 90°$) mit Kammabtastung (senkrechte Linien), passend zur Grundfrequenz der zugrundeliegenden Signale. Der Abstand zwischen den Linien $\nu_0(f_0)$ codiert die Wellenlänge der Grundfrequenz in Samples. Die Anzahl von interessierten Korrelationsmaxima ist mit $P = 2$ festgelegt. Die Abtastfrequenz beträgt $f_s = 24\,\text{kHz}$, Blocklänge $N = 2048$ **(a)** Korrelation eines Sägezahnsignals mit $f_0 = 150\,\text{Hz}$ **(b)** Korrelation des Sprachsignals mit $f_0 = 164\,\text{Hz}$.

Zwei beispielhaft aus der parametrischen Abtastung über Gleichung 3.1 generierte Ergebnisebenen sind in Abbildung 3.4 gezeigt. Bei den Quellen handelt es sich

jeweils um ein störungsfreies (kein Rauschen und kein Nachhall) Sägezahnsignal ($f_0 = 150\,\mathrm{Hz}$) bzw. Sprachsignal ($f_0 = 164\,\mathrm{Hz}$) aus frontaler Richtung ($\varphi_0 = 90°$). Die Ergebnisebene ist jeweils bei der wahren Einfallsrichtung bzw. Grundfrequenz nochmals als Querschnitt abgebildet. Es sind deutliche Mehrdeutigkeiten bei der Schätzung der Grundfrequenz und ebenso eine breite Schätzung in Bezug auf die Einfallsrichtung zu erkennen. Um eine robustere und genauere kombinierte Direction of Arrival (DoA) und Grundfrequenzschätzung zu generieren, wurden im Rahmen dieser Arbeit zusätzliche Erweiterungen implementiert und untersucht.



(a) Ungestörtes Sägezahnsignal $f_0 = 150\mathrm{Hz}$ aus $\varphi_0 = 90°$.

(b) Ungestörtes Sprachsignal $f_0 = 164\mathrm{Hz}$ aus $\varphi_0 = 90°$.

Abbildung 3.4: Ergebnisebenen $\rho'_{il}(\varphi_0, f_0)$ der kombinierten Grundfrequenz und Richtungsschätzung auf Grundlage der KKF zweier Eingangskanäle für **(a)** ungestörtes Sägezahnsignal mit $f_0 = 150\mathrm{Hz}$ und **(b)** Sprachsignal mit $f_0 = 164\mathrm{Hz}$ bei einer frontalen Einfallsrichtung $\varphi_0 = 90°$, $N = 2048\,\mathrm{Samples}$.

3.3 Kombinierte Schätzung über das Leistungsdichtespektrum

Wie in Abschnitt 2.3.5 beschrieben und auch in Abbildung 3.1 ersichtlich, bedingt eine effektive KKF-Ermittlung eine vorhergehende KLDS-Berechnung. Aber bereits mir Hilfe der KLDS kann eine kombinierte Schätzung erfolgen. Mit Gleichung (3.5) kann eine parametrische Abtastung des KLDS durchgeführt werden [KWH07]. Auf diesem Weg kann abermals eine über Grundfrequenz f_0 und Winkel φ aufgespannte Ergebnisebene bestimmt werden. Die dazu notwendige Berechnung gestaltet sich folgendermaßen:

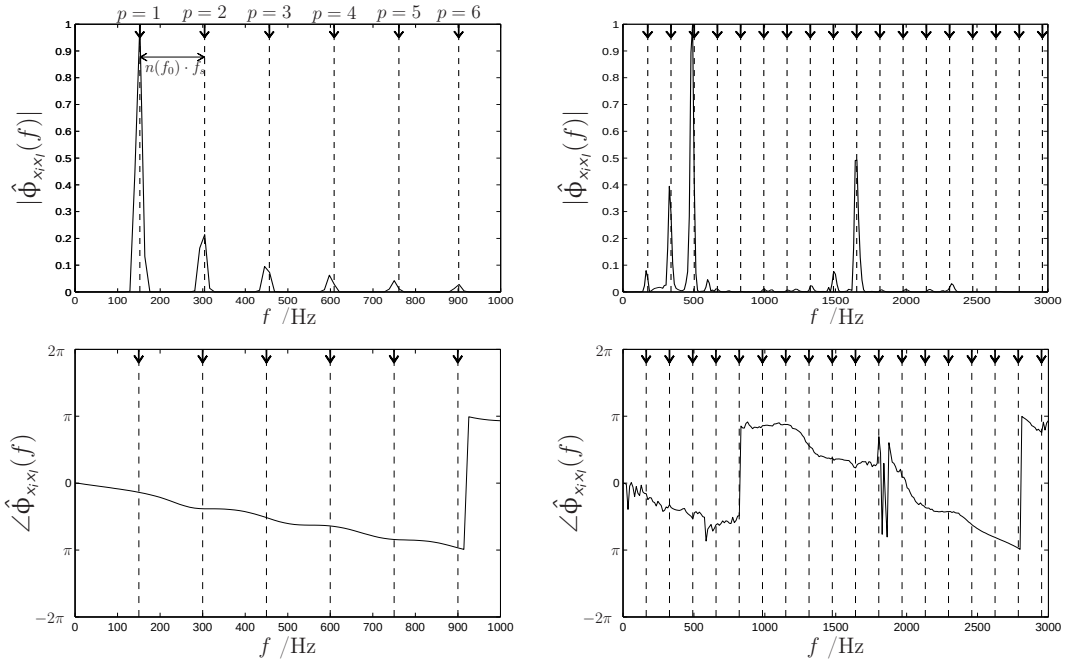
$$\tilde{\rho}_{il}(\varphi, f_0) = \sum_{n_p} |\hat{\phi}_{x_i x_l}[n_p]| \cdot \mathrm{T}\left\{\psi_{\mathrm{diff},p}(\psi, \varphi, f_0)\right\}, \tag{3.5}$$

$$\psi_{\mathrm{diff},p}(\psi, \varphi, f_0) = \underbrace{\angle\hat{\phi}_{x_i x_l}[n_p]}_{\psi_p} + p \cdot \underbrace{\kappa_{il}(\varphi)\Omega(f_0)}_{\psi_0 = 2\pi \frac{d_{il}\cos(\varphi) f_0}{c}}, , \tag{3.6}$$

$$n_p = p \cdot n(f_0) \qquad , \text{mit } p = 1 \ldots P, , \tag{3.7}$$

$$\tilde{\rho}'_{il}(\varphi, f_0) = |\tilde{\rho}_{il}(\varphi, f_0)| / \tilde{\rho}_{il,\max}(\varphi, f_0). \tag{3.8}$$

Der Term $n(f_0)$ definiert den Abstand der Harmonischen Vielfachen der Grundfrequenz f_0 im Spektrum. Weiterhin beschreibt die Phase ψ_0 die bei der Grundfrequenz f_0 und Einfallswinkel φ_0 erwartete Phasenverschiebung, wohingegen ψ_p die gemessene Phase ψ bei der p-fachen Harmonischen der Grundfrequenz f_0 beschreibt. Über den Betrag $|\hat{\phi}_{x_i x_l}[n]|$ des KLDS kann die harmonische Struktur des aufgenommenen Sprachsignals ermittelt werden. Die Zeitverzögerung des Eingangssignals zwischen zwei Mikrofonsignalen wird hingegen durch den Phasengang $\angle\hat{\phi}_{x_i x_l}[n]$ des KLDS codiert. Wann in Gleichung (3.5) ein auf eine Quelle schließendes Maximum entsteht, soll im Folgenden erläutert werden.



(a) KLDS mit Sägezahn **(b)** KLDS mit Sprachsignal

Abbildung 3.5: Betrags- und Phasengang des KLDS für ein **(a)** ungestörtes Sägezahnsignal mit $f_0 = 150\,\mathrm{Hz}$ und **(b)** ein ungestörtes Sprachsignal mit $f_0 = 164\,\mathrm{Hz}$. Der Einfallswinkel beträgt jeweils 34°. Die senkrechten Linien stellen die Abtastpunkte gemäß Gleichung (3.7) dar. Der Abstand zwischen den Linien entspricht der Grundfrequenz.

Der Betragsgang des KLDS wird an denen der betrachteten Grundfrequenz entsprechenden Abtastpunkte und deren Vielfachen durchwandert (vgl. Gleichung (3.7)). Es ist davon auszugehen, dass die Energie der interessierenden Nutzschallquelle deutlich größer ist als die der möglichen Störeinflüsse. Im Gegensatz zu Verfahren mit spektralem „Whitening“ nutzt diese Methode die Sammlung von Informationen für unabhängige Frequenzen [WK07]. Das Prinzip der Abtastung des KLDS ist in Abbildung 3.5 illustriert.

Phasentransformation

Die ermittelten Beträge $|\hat{\Phi}_{x_i x_l}[n_p]|$ der Kammabtastung werden zudem mit einer Phasentransformationsfunktion $\mathrm{T}\{\cdot\}$ multipliziert. Deren Aufgabe ist es, eine hohe Gewichtung der Beträge zu gewährleisten, wenn die erwartete Phase ψ_0 mit der wahren Phase ψ übereinstimmt. Bei Phasenübereinstimmung ergibt sich nach Gleichung (3.6) ein Betrag von Null bzw. durch allgemeine Mehrdeutigkeiten der Phase ein beliebiges Vielfaches von 2π [WK07]. Als Transformationsfunktion $\mathrm{T}\{\cdot\}$ kann dementsprechend prinzipiell jede reellwertige, 2π periodische, gerade Funktion eingesetzt werden. Diese Vorraussetzung erfüllend wurden die Transformationsfunktionen $\mathrm{T}_1\{\cdot\}$, $\mathrm{T}_2\{\cdot\}$ aus [WK07] und der eigene Entwurf $\mathrm{T}_3\{\cdot\}$ untersucht. Die Gewichtung beruht jeweils auf der Differenz der gemessenen Phase ψ mit der bei der parametrischen Abtastung aktuell erwarteten Phase ψ_0 für eine bestimmte Grundfrequenz aus einer bestimmten Richtung (vgl. Gleichung (3.6)).

$$\mathrm{T}_1\{\psi_{\mathrm{diff},p}(\psi,\varphi,f_0)\} = \cos(\psi_{\mathrm{diff},p}(\psi,\varphi,f_0)) \tag{3.9}$$

$$\mathrm{T}_2\{\psi_{\mathrm{diff},p}(\psi,\varphi,f_0)\} = \frac{1}{1+\beta-\cos(\psi_{\mathrm{diff},p}(\psi,\varphi,f_0))} \tag{3.10}$$

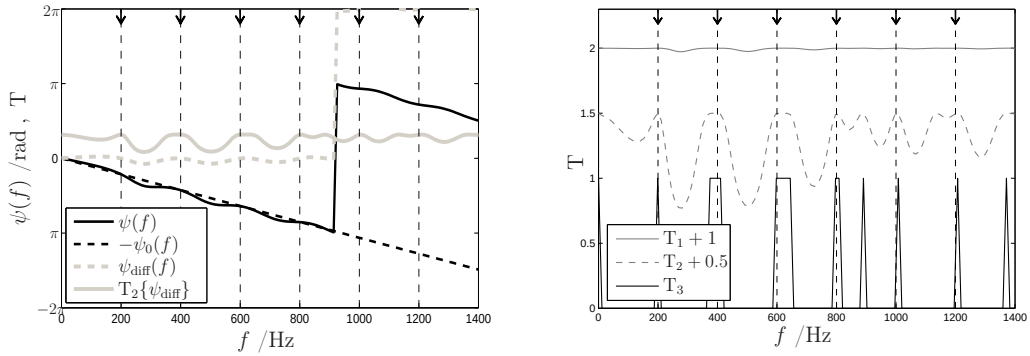
$$\mathrm{T}_3\{\psi_{\mathrm{diff},p}(\psi,\varphi,f_0)\} = \begin{cases} 1 & \text{, wenn } (\psi_{\mathrm{diff},p}(\psi,\varphi,f_0) \bmod 2\pi) < \beta \\ 0 & \text{, sonst} \end{cases}, \tag{3.11}$$

$$\text{mit} \quad (a \mod m) := a - m \cdot (a \operatorname{div} m) \tag{3.12}$$

Dabei bezeichnet $a \operatorname{div} m$ den zur Null gerundeten Quotienten $a \,/\, m$ (symmetrische Moduloberechnung).

Die Funktionsweise der Phasentransformation ist in Abbildung 3.6a wiedergegeben. Eine aus der Summe der gemessenen Phase $\psi(f)$ und der berechneten Phase $\psi_0(f)$ resultierende Funktion wird an den durch die Kammabtastung relevanten Stellen mit der Phasentransformation $\mathrm{T}\{\cdot\}$ zu einem Gewichtungsfaktor. Wenn, wie in Abbildung 3.6a gegeben, die theoretischen und realen Phasen an den betrachteten Abtastpunkten übereinstimmen, so ergibt sich aus der Summe ein Wert nahe Null oder nahe $2\pi \cdot p$ ($p \in \mathbb{N}$) und dementsprechend eine hohe Gewichtung aus der Phasentransformation. Dabei übernimmt diese gleichzeitig die Funktion einer Phasenkorrektur (engl. wrapping/unwrapping), welche beim Überschreiten der 2π Phasengrenze in der gemessenen Phase auftritt.

Die mit Gleichungen (3.9), (3.10) und (3.11) vorgestellten Transformationsfunktionen können in Abbildung 3.6b exemplarisch verglichen werden. Das zugrunde liegende Szenario ist mit dem aus Abbildung 3.6a identisch. Beim Vergleich der Graphen lässt sich feststellen, dass die Variante $\mathrm{T}_1\{\cdot\}$ bei richtiger Winkelschätzung (wie in der Abbildung vorliegend) kaum Einfluss auf das Ergebnis nimmt. Wohingegen bei den Versionen $\mathrm{T}_2\{\cdot\}$ und $\mathrm{T}_3\{\cdot\}$ eine klarer Einfluss auszumachen ist. Es ist zu erwähnen, dass besonders die Funktion $\mathrm{T}_3\{\cdot\}$ anfälliger auf Parameterveränderungen reagiert, was das Wählen des Schwellwertes β betrifft. Die Funktion $\mathrm{T}_3\{\cdot\}$ ist dadurch sehr sensibel gegenüber Winkeländerungen. Bereits kleine Winkelunterschiede ($< 1°$) können großen Einfluss auf das Ergebnis haben. Bei einer Winkelabtastung von 1°, sowie der prinzipbedingten Abhängigkeit der Winkelgenauigkeit vom Einfallswinkel führt diese nicht immer zum optimal möglichen Resultat.

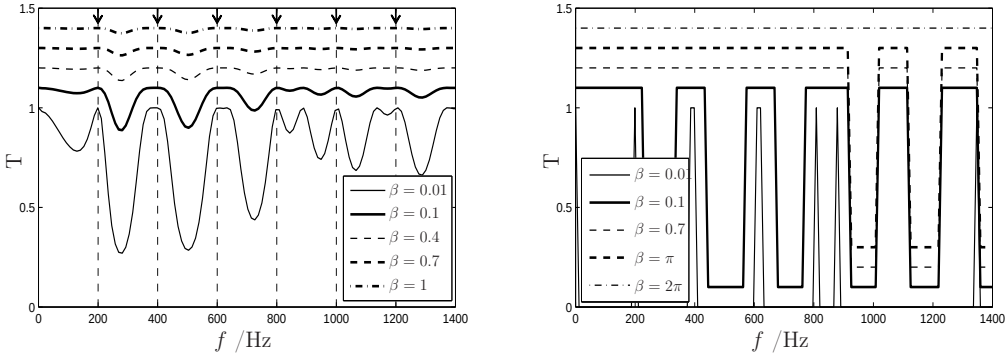


(a) **(b)**

Abbildung 3.6: **(a)** Phasengang $\psi(f)$ eines ungestörten Sägezahnsignals ($f_0 = 200\,\mathrm{Hz}$) und dessen Phasentransformation. Die gemessene Phase $\psi(f)$ stammt von einem Kurzzeit-KLDS (85 ms). Zwei Mikrofonen bei einer Einfallsrichtung von 34°. Die angenommene Phase ψ_0 entspricht ebenso einem Einfallswinkel von 34°. $\psi_{\mathrm{diff}}(f)$ ist die Summe von $\psi_0(f)$ und $\psi(f)$. Die Phasentransformation $\mathrm{T}_2\{\psi_{\mathrm{diff}}\}$ gewährleistet neben der Phasenkorrektur (vgl. 900 Hz) bei Übereinstimmung der Messung mit der Erwartung einen hohen Gewichtungsfaktor bei den Kammabtastpunkten (senkrechte Linien). **(b)** Resultierende Gewichtungsfunktionen entsprechend dem Beispiel aus Abbildung (a) für alle drei Transformationsfunktionen. Der Schwellwert lag jeweils bei $\beta = 0,01$. Die Varianten $\mathrm{T}_1\{\cdot\}$ und $\mathrm{T}_2\{\cdot\}$ wurden zur besseren Visualisierung mit einem Offset belegt.

In Abbildung 3.7 sind weitere Phasentransformationsergebnisse abgebildet. Das Szenario entspricht abermals dem aus Abbildung 3.6, jedoch wurden hier nochmals unterschiedliche β-Paramter verwendet. Mit wachsenden β tendiert die Phasengewichtung $\mathrm{T}_2\{\cdot\}$ aus Abbildung 3.7a erwartungsgemäß zur gleichen Gewichtungsfunktion, wie sie die $\mathrm{T}_1\{\cdot\}$ Variante in Abbildung 3.6b aufweist. Der Charakter der Funktion wird zunehmend undifferenzierter und besitzt bei $\beta = 1$ einen

nahezu konstanten Verlauf. Auch die selbst entworfene Variante $\mathrm{T}_3\{\cdot\}$ zeigt mit wachsenden Gewichtungsparameter β einen monotonen Verlauf.



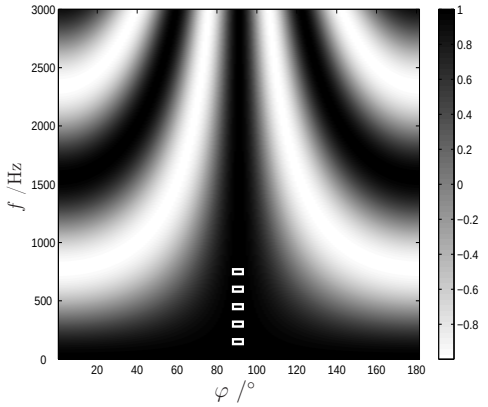
(a) $\mathrm{T}_2\{\cdot\}$ mit unterschiedlichen β. Zusätzliches Offset erhört sich jeweils um 0,1, beginnend mit 0.

(b) $\mathrm{T}_3\{\cdot\}$ mit unterschiedlichen β. Zusätzliches Offset erhört sich jeweils um 0,1, beginnend mit 0.

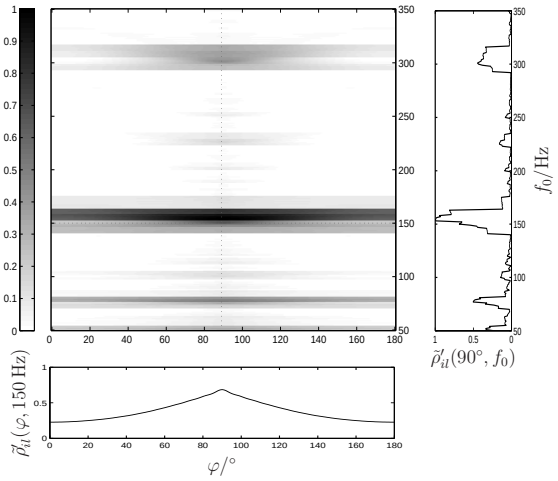
Abbildung 3.7: Weitere Resultate der Phasentransformation mit verschiedenen β-Parameter, entsprechend dem Szenario aus Abbildung 3.6 sind in **(a)** für die Variante $\mathrm{T}_2\{\cdot\}$ und in **(b)** für $\mathrm{T}_3\{\cdot\}$ abgebildet. Zur besseren Visualisierung sind die verschiedenen Kurven aufsteigend mit einem Offset von 0 bis 0,4 beginnden bei der Kurve für $\beta = 0.01$ belegt.

Die durch die Transformationsfunktionen resultierenden Gewichtungspattern sind in Abbildung 3.8 für eine Vorzugsrichtung von $\varphi_0 = 90°$ dargestellt (linke Seite). Für die Grundfrequenz $f_0 = 150\,\mathrm{Hz}$ liegen die ersten $P = 5$ Kammabtastwerte an den durch die Rechtecke gekennzeichneten Positionen. Es ist zu erkennen, dass bei der Verwendung von $\mathrm{T}_1\{\cdot\}$ durch die breite Vorzugsrichtung eine breite Winkelschätzung resultiert. Das räumliche Aliasing erzeugt weitere Nebenkeulen, die zu verfälschenden Nebenmaxima in der Ergebnisebene führen. Die Transformationsfunktionen $\mathrm{T}_2\{\cdot\}$ und $\mathrm{T}_3\{\cdot\}$ versuchen diese Effekte zu minimieren, wobei durch den Faktor β ein direkter Einfluss auf die breite der Vorzugsrichtung genommen werden kann. Transformationsfunktion $\mathrm{T}_1\{\cdot\}$ sowie $\mathrm{T}_2\{\cdot\}$ sind in [WK07] vorgeschlagen, wohingegen Variante $\mathrm{T}_3\{\cdot\}$ eine auf den Minimalanforderungen für die Transformationsfunktion aufbauend, selbsterstelle Version ist. Der Vergleich der drei dargestellten Phasengewichtungen aus Abbildung 3.8 (a),(c) und (e) verdeutlicht die unterschiedlichen Gewichtungen in ihrer Richtungscharakteristik.

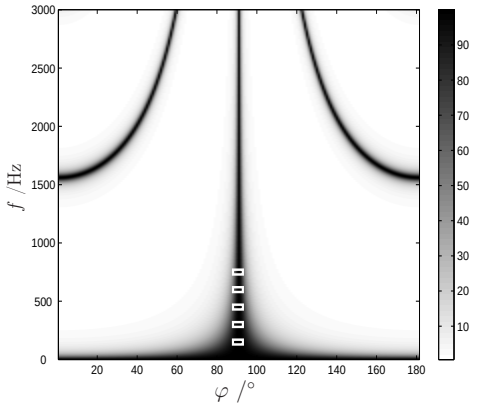
Beispielhafte, mit den verschiedenen Gewichtungsfunktionen berechnete Ergebnisebenen sind auf der rechten Seite von Abbildung 3.8 zu sehen. Bei dem Testsignal handelt es sich jeweils um ein störungsfreies Sägezahnsignal ($f_0 = 150\,\mathrm{Hz}$) aus frontaler Einfallsrichtung. Bei dem Vergleich der Abbildungen ist besonders die sich verbessernde Richtungsschätzung zu nennen. Im Bereich des interessierenden Einfallswinkels treten jedoch weiterhin Fehlschätzungen der Grundfrequenz auf, was beim kontrollieren der Querschnittsdiagramme für die erwartete Einfallrichtung und Grundfrequenz auszumachen ist.



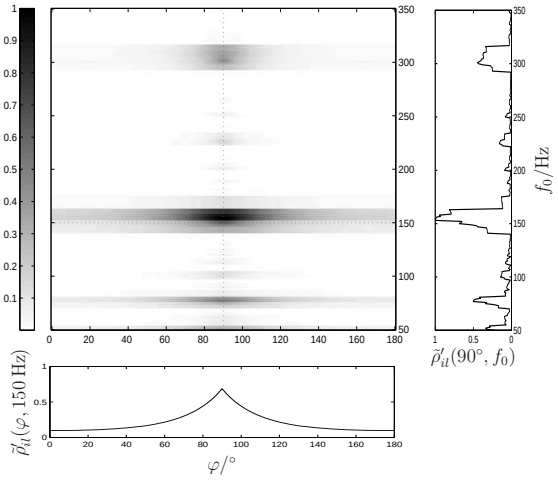
(a) $\cos(\varphi, f)$



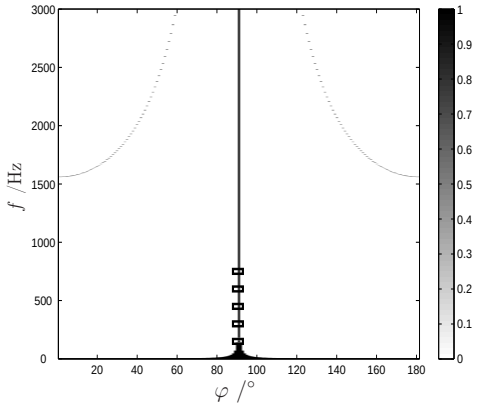
(b) Gleichung (3.5) mit $\mathrm{T}_1\{\cdot\}$



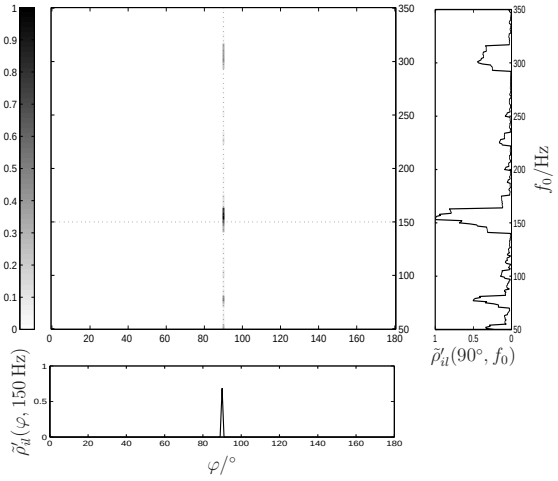
(c) $\frac{1}{1+\beta-\cos(\varphi,f)}$ mit $\beta = 0.01$



(d) Gleichung (3.5) mit $\mathrm{T}_2\{\cdot\}$



(e) $((\varphi, f) \mod 2\pi)$ mit $\beta = 0.01$



(f) Gleichung (3.5) mit $\mathrm{T}_3\{\cdot\}$

Abbildung 3.8: Linke Seite: Gewichtungspattern für Phasentransformation $\mathrm{T}_{1,2,3}\{\cdot\}$ für $\varphi_0 = 90°$. Rechtecke sind Abtastpunkte n_p mit $p = 1 \ldots 5$. Rechte Seite: Ergebnisebenen mit ungestörten Sägezahn $f_0 = 150\,\mathrm{Hz}$, $\varphi_0 = 90°$, unter Verwendung der jeweiligen Transformationsfunktion. Blockgröße $N = 2048$, f_s=24 kHz.

3.4 Mehrkanalberechnung

Um die Effekte des räumlichen Aliasing (vgl. Abschnitt 3.3) zu reduzieren, wurden mehrere Mikrofonpaare zur Schätzung herangezogen und nicht wie bisher nur ein Mikrofonpaar verwendet. Um die Ergebnisse aller Mikrofonpaare zu einer einheitlichen Schätzung zusammenzufassen, wurden zwei unterschiedliche Ansätze verfolgt. Einerseits wurden die Resultate der jeweiligen Mikrofonpaare durch eine Mittelung nach Gleichung (3.13) fusioniert,

$$\rho(\varphi, f_0) = \frac{M^2 - M}{2} \sum_{i=1}^{M} \sum_{l=i+1}^{M} \rho_{il}(\varphi, f_0), \tag{3.13}$$

wobei M die Anzahl der Mikrofone definiert [WK07]. Andererseits wurde auch erstmals der Ansatz der Multichannel Cross-Correlation (MCCC) zur Fusion der einzelnen kombinierten Schätzungen implementiert. Dabei wurde die MCCC aus [CBH03, CBH06, CHB05, Mey08] für die hier beschriebene Anwendung adaptiert. Eigentliche Vorraussetzung für das MCCC-Verfahren ist die Bestimmbarkeit der Time Delay Estimation (TDE) zwischen den Mikrofonen der jeweiligen Mikrofonpaare. Das Verfahren wurde jedoch auf die zusätzliche Schätzung der Grundfrequenz erweitert. Durch diese selbst vorgenommene Anpassung ist es möglich, das dem MCCC-Verfahren zu Grunde liegende Prinzip auf die kombinierte Grundfrequenz- und Richtungsschätzung anzuwenden. Die Werte aller Mikrofonpaare werden während der parametrischen Abtastung in einer Korrelationsmatrix zusammengefasst:

$$\mathbf{P}(\varphi, f_0) = \begin{pmatrix} \rho_{11}(\varphi, f_0) & \rho_{12}(\varphi, f_0) & \cdots & \rho_{1l}(\varphi, f_0) \\ \rho_{21}(\varphi, f_0) & \rho_{22}(\varphi, f_0) & \cdots & \rho_{2l}(\varphi, f_0) \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{i1}(\varphi, f_0) & \rho_{i2}(\varphi, f_0) & \cdots & \rho_{il}(\varphi, f_0) \end{pmatrix}. \tag{3.14}$$

Die Elemente auf der Hauptdiagonalen von $\mathbf{P}$ können zu 1 gesetzt werden, wenn man zuvor eine Normierung der Matrixelemente mit $1/\sqrt{\rho_{ll}(\varphi, f_0)\rho_{ii}(\varphi, f_0)}$ sicherstellt. Da eine Vertauschung der Sensoren von $\rho_{il}(\varphi, f_0)$ zu $\rho_{li}(\varphi, f_0)$ mit dann notwendiger Invertierung der Time Delay Difference (TDD) zu einem identischen Ergebnis führt, kann die Matrix abermals vereinfacht werden. Die vereinfachte Schreibweise lautet anschließend:

$$\check{\mathbf{P}}(\varphi, f_0) = \begin{pmatrix} 1 & \rho_{12}(\varphi, f_0) & \cdots & \rho_{1l}(\varphi, f_0) \\ \rho_{12}(\varphi, f_0) & 1 & \cdots & \rho_{2l}(\varphi, f_0) \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1l}(\varphi, f_0) & \rho_{2l}(\varphi, f_0) & \cdots & 1 \end{pmatrix}. \tag{3.15}$$

Jedes Element der Ergebnisebene der kombinierten Schätzung errechnet sich anschließend aus der Determinanten der aufgestellten Matrix. Durch die Normierung

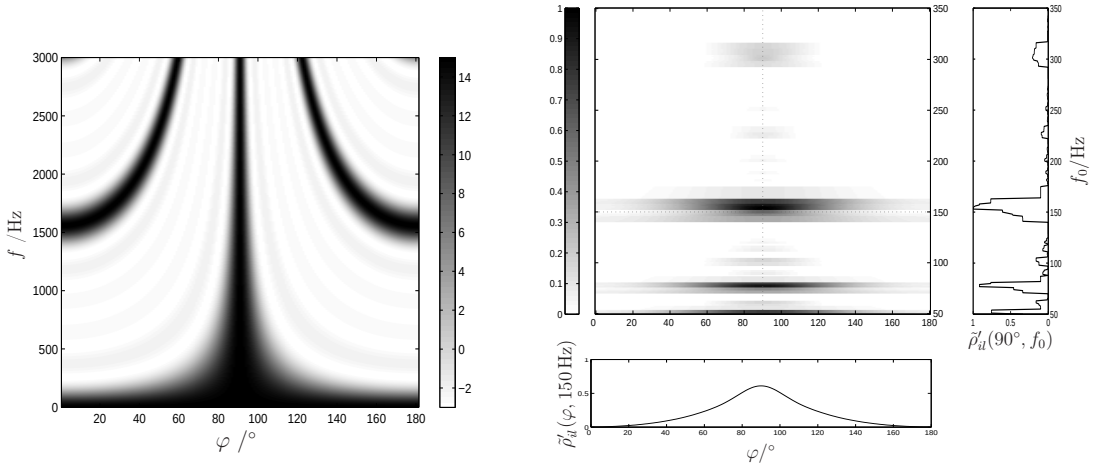
der jeweiligen Schätzung ρ ist nach [CBH03] gewährleistet, dass die Determinante einen Wert zwischen 0 und 1 annimmt,

$$0 < \det\left(\check{\mathbf{P}}(\varphi, f_0)\right) < 1. \tag{3.16}$$

Da bei einer hohen Übereinstimmung der Werte in der Matrix die Determinatne klein wird (Vektoren zeigen alle in die gleiche Richtung), ergeben sich die Elemente der kombinierten DoA-Grundfrequenz Ergebnisebene zu

$$\check{\rho}(\varphi, f_0) = 1 - \det\left(\check{\mathbf{P}}(\varphi, f_0)\right). \tag{3.17}$$

Der Vorteil des MCCC-Verfahrens, gegenüber der Berechnung der Mittelwerte nach Gleichung (3.13), liegt in der Bedeutung der Determinante. Diese kann als Volumen des von den Vektoren der Matrize aufgestellten merhdimensionalen Raumes angesehen werden. Nur wenn die Vektoren in die gleiche Richtung weisen, also ähnliche Ergebnisse vorweisen, wird das Volumen des virtuellen Raumes gering und die Schätzung des Einfallswinkels und der Grundfrequenz hoch. Bei schlechter Übereinstimmung der Schätzungen der einzelnen Mikrofonpaare, wird die Determinatne groß und das Ergebnis ist in der Summe eine geringe Schätzung. Der Vorteil gegenüber einer einfachen Mittelung ist, dass durch die Mehrdimensionalität sich die möglichen Abweichungen nicht zufällig ausgleichen und dementsprechend ein falsches Ergebnis schätzen.



(a) Phasengewichtung $\cos(\varphi, f)$ **(b)** Gleichung 3.5 mit T_1

Abbildung 3.9: **(a)** Mehrkanalberechnung der Richtungsgewichtung mit T_1, **(b)** Ergebnisebene der kombinierten Schätzung bei einem ungestörten Sägezahnsignal $f_0 = 150\,\text{Hz}, \varphi_0 = 90°$, durch Mittelung von 15 Mikrofonpaaren, $N = 2048\,\text{Samples}$, $f_s = 24\,\text{kHz}$.

Zur Veranschaulichung des Mehrwertes der Mehrkanalberechnung ist in Abbildung 3.9 nochmals eine Phasengewichtungsfunktion nach Gleichung (3.9) bei Ver-

wendung von 15 Mikrofonpaaren dargestellt. Anhand dieser Gewichtungsfunktion ist der, das räumliche Aliasing minimierende, Effekt der Mehrkanalberechnung auszumachen. Das sich ergebende Resultat ist auf der rechten Seite abgebildet. Bei einem Vergleich mit Abbildung 3.8a,b ist die verbesserte räumliche Schätzung erkennbar.

3.5 Cepstrum-Gewichtung

In diesem Abschnitt soll eine auf dem Cepstrum beruhende Modifikation beschrieben werden, die eine zusätzliche Hervorhebung der Grundfrequenz bewirkt. Das Verfahren ist aus [HKO08] entnommen. Eine einleitende Erläuterung zur Interpretation des Cepstrums und zur Grundfrequenzerkennung durch das Cepstrum ist in Abschnitt 2.2.3 zu finden. Die hier vorgestellte Gewichtung $w[k]$ der Ergebnisebene der kombinierten Schätzung beruht auf einer halbwellengleichgerichteten cepstralen Repräsentation der KKF,

$$w_{il}[k] = |IDFT(\log_{10}(|DFT(\hat{r}^{+}_{x_i x_l}[k])| + \underbrace{1e^{-6}}_{\varepsilon}|)), \tag{3.18}$$

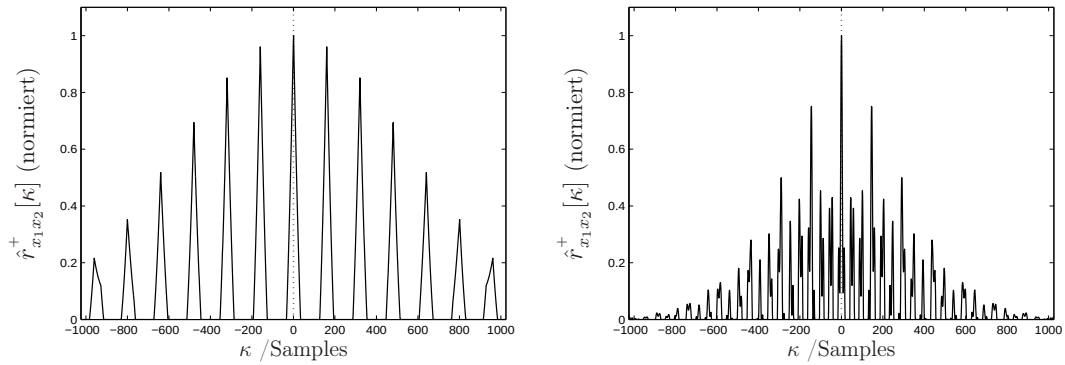
$$w_{il,mov}[k] = \max(w_{il}[k-2]...w_{il}[k+2]). \tag{3.19}$$

Die Halbwellengleichrichtung[1] der KKF soll die Suche nach Grundfrequenzen unterstüzen, die sich in den Abständen der Korrelationsmaxima codieren, jedoch von Störeinflüssen überlagert sein können. Daher kann durch Vernachlässigung der negativen Werte der KKF die cepstrale Gewichtung optimiert werden. Dem KLDS wird ein Versatz $\varepsilon = 1e^{-6}$ aufaddiert, um den Logarithmus einzugrenzen.

Wie aus Gleichung (3.18) zu entnehmen ist, interessiert nur der Betrag der Gewichtungsbildung. Die Phase der Cepstrumsrepräsentation enthält keine Informationen über die Grundfrequenzen. Da somit die Phase aller Frequenzkomponenten zu Null gesetzt ist, kann die Gewichtung für alle Einfallsrichtungen verwendet werden.

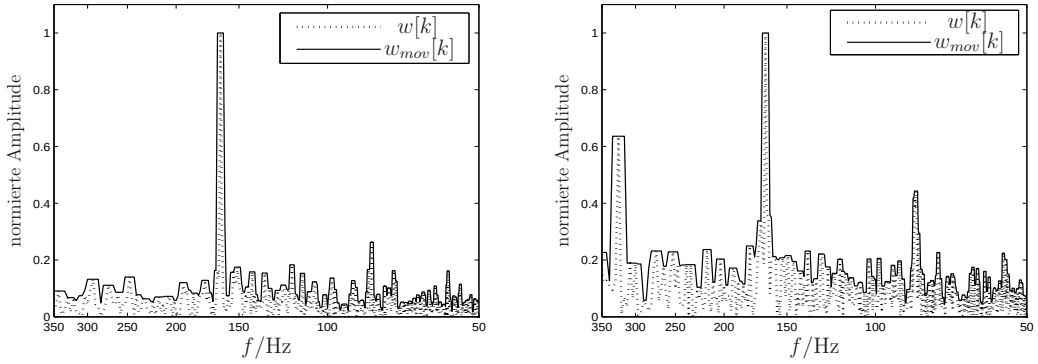
Die halbwellengleichgerichteten Kurzzeitkreuzkorrelationen eines ungestörten Sägezahnsignals sowie eines Sprachsignals sind in Abbildung 3.10 zu sehen. Ferner wird eine Glättungsfunktion mit Gleichung (3.19) eingeführt. Mit Hilfe dieses fließenden Maximums werden die Abtastwerte der Gewichtung konstanter und weniger schwankend [HKO08]. Eine resultierende Gewichtungsfunktion nach Gleichung (3.18) und (3.19) ist für ein ungestörtes Sprachsignal mit $f_0 = 164\,\mathrm{Hz}$ in Abbildung 3.11a dargestellt. Abbildung 3.11b zeigt die resultierende Gewichtung ohne Halbwellengleichrichtung. Es ist sowohl die Glättung durch Gleichung (3.19) in beiden Abbildungen als auch der Gewinn der Halbwellengleichrichtung bei Vergleich der beiden Abbildungen erkennbar.

[1] Die Notation $\cdot^{+}$ definiert eine Halbwellengleichrichtung.



(a) $\hat{r}^{+}_{x_1x_2}[\kappa]$ von einem Sägezahnsignal. **(b)** $\hat{r}^{+}_{x_1x_2}[\kappa]$ von einem Sprachsignal.

Abbildung 3.10: Halbwellengleichgerichtete Kurzzeit-KKF $\hat{r}^{+}_{12}[k]$ (85 ms) eines **(a)** ungestörten Sägezahnsignals $f_0 = 150\,\text{Hz}$ und **(b)** ungestörten Sprachsignals $f_0 = 164\,\text{Hz}$. Aufgenommen durch ein Mikrofonpaar bei einer Einfallsrichtung von $\varphi_0 = 90°$.



(a) mit Halbwellengleichrichtung **(b)** ohne Halbwellengleichrichtung

Abbildung 3.11: Cepstrumsgewichtung eines Kurzzeitsprachsignals (85 ms) mit Grundfrequenz $f_0 = 164\,\text{Hz}$, wobei **(a)** mit und **(b)** ohne Halbwellengleichrichtung der KKF vorgenommen wurde. Die durchgezogene Linie stellt die geglättete Gewichung $w_{mov}[k]$ der gestrichelten Linie dar.

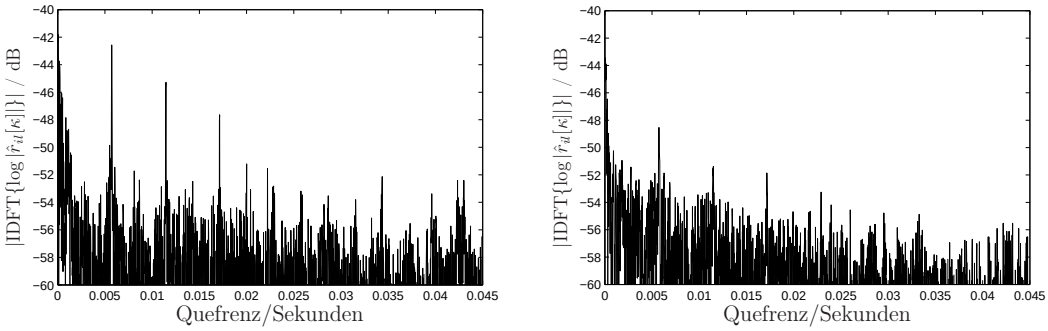
Die Gewichtung $w_{mov}[k]$ kann auf die einzelen Ergebnisfelder pro Mikrofonpaar separat oder nach der Fusion auf das resultierende Ergebnisfeld angewandt werden. Dies ist möglich, da die cepstrale Gewichtung lediglich zur Verbesserung der Grundfrequenzschätzung dient und diese nicht vom Mikrofonabstand abhängig ist. Mit Gleichung (3.20) wird die Gewichtung jeweils für ein Mikrofonpaar auf die entsprechende Ergebnisebene angewandt,

$$\rho_{\text{cep},il}(\varphi, f_0) = w_{mov,il}[\lfloor \nu_0(f_0) \rfloor] \cdot \rho_{il}(\varphi, f_0). \tag{3.20}$$

Wobei dies unter Berücksichtigung der untersuchten Grundfrequenzen mit Hilfe einer Look-up Tabelle nach $\nu_0(f_0)$ erfolgt, um eine Zuordnung der Quefrenzen

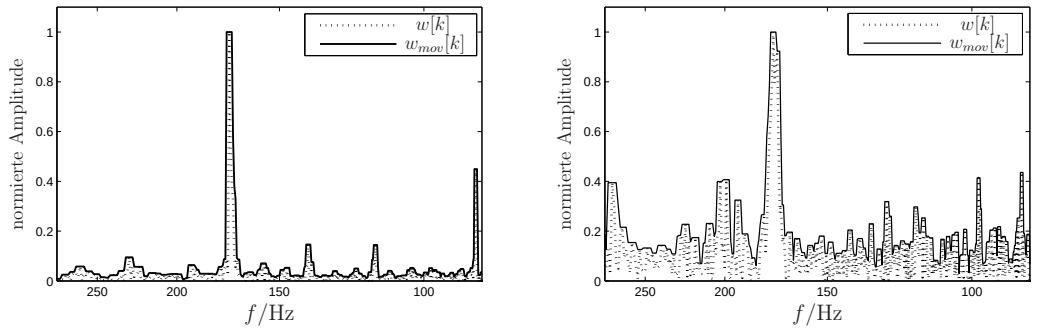
des Cepstrums zu den Grundfrequenzen zu gewährleisten. Eine Kombination der Ergebnisebenen erfolgt dann im Anschluss gemäß Abschnitt 3.4.

Abbildung 3.12 verdeutlicht zwei Cepstrumsgewichtungen ausgehend von unterschiedlichen akustischen Bedingungen. Zum einen diente ein unverfälschtes Nutzsignal als Grundlage, zum anderen wurde das Signal mit eine Rauschen von 0 dB SNR und einer Nachhallzeit von $\tau_{60} \approx 550\,\text{ms}$ überlagert. Es ist auch bei der zweiten mit Störeinflüssen überlagerte Version noch eine Hervorhebung des Nutzsignals von den Störeinflüssen erkennbar. So dass eine Anwendung der Cepstrumsgewichtung auch bei schlechten akustischen Bedingungen noch ihre Berechtigung hat.



(a) Cepstrum bei SNR $\to \infty$ und keinem Nachhall $\tau_{60} \approx 0\,\text{ms}$

(b) Cepstrum bei 0 dB SNR und realer RIR mit $\tau_{60} \approx 500\,\text{ms}$



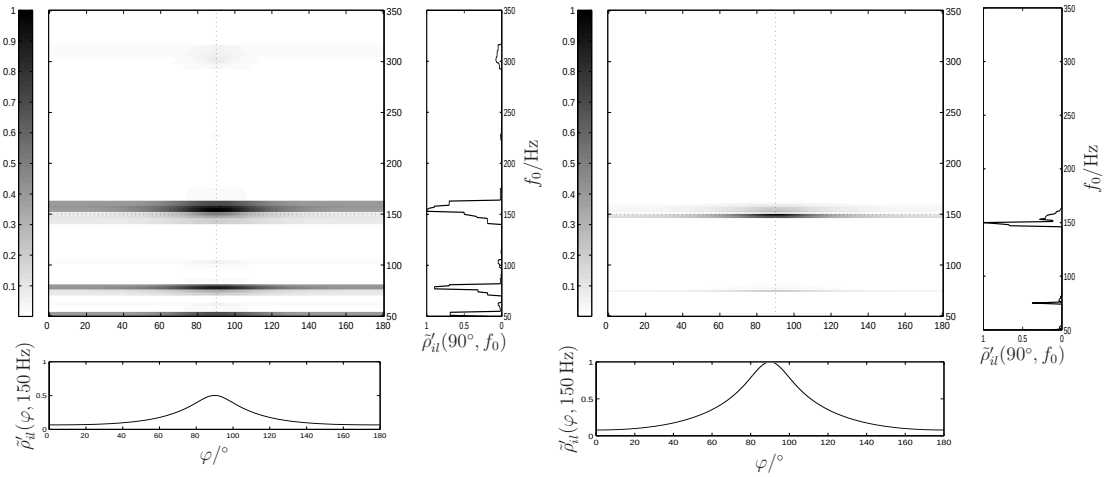
(c) mit Gleichung (3.19) aus **(a)** resultierende Cepstrum-Gewichtung $w_{il,mov}[k]$

(d) mit Gleichung (3.19) aus **(b)** resultierende Cepstrum-Gewichtung $w_{il,mov}[k]$

Abbildung 3.12: Cepstrum **(a,b)** und Cepstrumssgewichtung **(c,d)** eines Kurzzeitsprachsignals (85 ms) mit Grundfrequenz $f_0 = 164\,\text{Hz}$ ($f_s = 48\,\text{kHz}$). Die linke Seite wurde ohne Störgeräusch und Nachhall berechnet, wohingegen auf der rechten Seite das Nutzsignal mit 0 dB SNR und realen RIR ($\tau_{60} \approx 550\,\text{ms}$)überlagert wurde.

In Abbildung 3.13 ist die Auswirkung der Cepstrumsgewichtung auf das Resultat der kombinierten Grundfrequenz und Richtungsschätzung für ein ungestörtes Sägezahnsignal dargestellt. Beide Graphen stellen die kombinierte Schätzung einer identischen akustischen Szene dar. Es wurde jeweils die kombinierte Schätzung auf Grundlage des KLDS mit 15 Mikrofonpaaren, der MCCC-Kombination und der

$\mathrm{T}_1\{\cdot\}$ Phasengewichtungsfunktion durchgeführt. Ein Vergleich der beiden Darstellungen zeigt, dass durch die Cepstrum-Gewichtung die wahre Grundfrequenz der Quelle besser geschätz wird, sowie deren Harmonische stärker unterdrückt werden.



(a) ohne Cepstrum-Gewichtung **(b)** mit Cepstrum-Gewichtung

Abbildung 3.13: Ergebnisebenen der kombinierten Schätzung auf Grundlage der KLDS, sechs Eingangskanälen und MCCC-Kombination, **(a)** ohne Cepstrumsgewichtung, **(b)** mit Cepstrumsgewichtung, Blockgröße $N = 2048$ Sample. Es handelt sich jeweils um ein Sägezahnsignal $f_0 = 150$Hz bei einer frontalen Einfallsrichtung $\varphi_0 = 90°$.

3.6 Filterbankvorverarbeitung

Bis zu diesem Abschnitt wurde bei den Betrachtungen immer von einer einzelnen zu detektierenden Quelle ausgegangen. Es ist jedoch Ziel dieser Arbeit die kombinierte Grundfrequenz- und Richtungsschätzung für simultane Mehrsprechersituationen zu untersuchen. Bei dem Versuch mit den bisher beschriebenen Erweiterungen zwei simultane Sprecher zu erkennen, tendiert die Schätzung entweder dazu nur eine dominate Quelle darzustellen oder die Einfallsrichtungen und Grundfrequenzen entsprechen nicht den reellen Gegebenheiten. Um diesen Mangel entgegenzuwirken wurde in [KOH08] eine weitere Vorverarbeitung vorgestellt. Inspiriert durch das auditorische Modell des Menschen wurde in [KOH08] eine F-bandige Bandpass-Gammatonefiltervorverarbeitung der FFT-Spektren eingeführt. Vor der Erzeugung der KKF bzw. KLDS pro Mikrofonpaar werden nun alle Mikrofonsignale zuerst mit einer Gammatonfilterbank gefiltert. In dieser Arbeit wurde die Gammatonfilterung im Frequenzbereich vollzogen, wodurch sich die Filterung durch eine Fensterung im Spektralbereich realisieren lässt,

$$\mathsf{x}_{ig_p}[n] = \mathsf{x}_i[n] \cdot \mathrm{g}_p[n]. \tag{3.21}$$

Die FFT transformierten Mikrofonsignale $\mathsf{x}_i[n]$ werden mit den Bandpassgewichtungsfunktionen $\mathrm{g}_p[n]$ multipliziert. Zur Erzeugung der Gewichtungsfunktionen $\mathrm{g}_p[n]$ wurde die Toolbox von [Ell09] verwendet. Diese enthält die Funktion `fft2gammatonemx`, welche eine Matrix mit Gewichtungen erzeugt (reellwertige FIR-Filter), um ein Spektrum in mehrere gammatonegewichtete Spektren aufzuteilen.

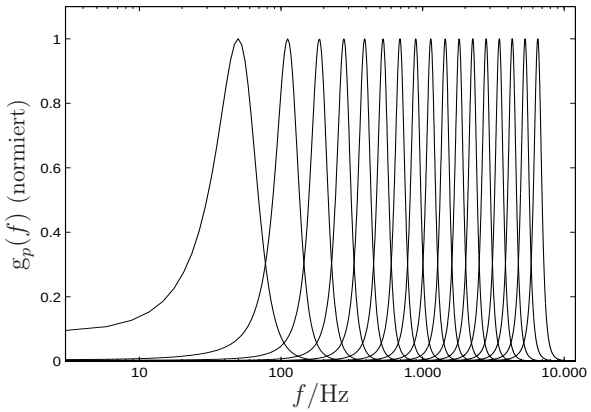


Abbildung 3.14: 17 Bandpass-Gammatonfenster mit 50 Hz bis 8000 Hz Mittenfrequenz (gleichverteilt auf der ERB-Skala).

Aus diesen bandbegrenzten Spektren $\mathsf{x}_{i\mathrm{g}_p}[n]$ werden die KKF bzw. die KLDS für jedes Filterband getrennt berechnet. Die bandbegrenzten Resultate $\hat{r}_{il,\mathrm{g}_p}[\kappa]$ werden normiert und zu einer gesammten KLDS oder KKF aufsummiert:

$$\hat{r}_{il,\mathrm{g}}[\kappa] = \frac{1}{F}\sum_{p=1}^{F}\frac{\hat{r}_{il,\mathrm{g}_p}[\kappa]}{\max(\hat{r}_{il,\mathrm{g}_p}[1...\kappa...N])}, \tag{3.22}$$

$$\hat{\phi}_{il,\mathrm{g}}[n] = \frac{1}{F}\sum_{p=1}^{F}\frac{\hat{\phi}_{il,\mathrm{g}_p}[n]}{\max\left(\left|\hat{\phi}_{il,\mathrm{g}_p}[1...n...\frac{N}{2}+1]\right|\right)}. \tag{3.23}$$

Die Normierung auf die Maximalwerte innerhalb der bandbegrenzten Spektren verursacht einen höheren Einfluss der Maxima bezogen auf ein nicht unterteieltes Spektrum. Somit erhalten die Nebenmaxima durch Harmonische in denen von der Grundfrequenz unbeeinflussten, bandbegrenzten Spektren eine höhere Bedeutung.

Zur Verdeutlichung der Arbeitschritte ist in Abbildung ein zugehöriges Flussdiagramm dargestellt. In der späteren Anwendung wurden $F = 64$ überlappende Bänder verwendet (ähnlich [HR10]), deren Mittenfrequenzen gleichmäßig auf die ERB-Skale von 50 Hz bis 8000 Hz verteilt sind. Der Mehrgewinn dieser bandbegrenzten Vorverarbeitung liegt in dem unterschiedlichen Informationsgehalt der einzelnen bandpassgewichteten KKF bzw. KLDS. Die tiefen Gammatonbandpässe (< 1 kHz) heben vor allem Pitch Informationen hervor, wobei die höheren Frequenzbänder die DoA-Schätzung hervorheben [KOH08]. Durch die Bandpassfilterung soll es nach [KOH08] besser möglich sein, mehrere simultane Sprecher

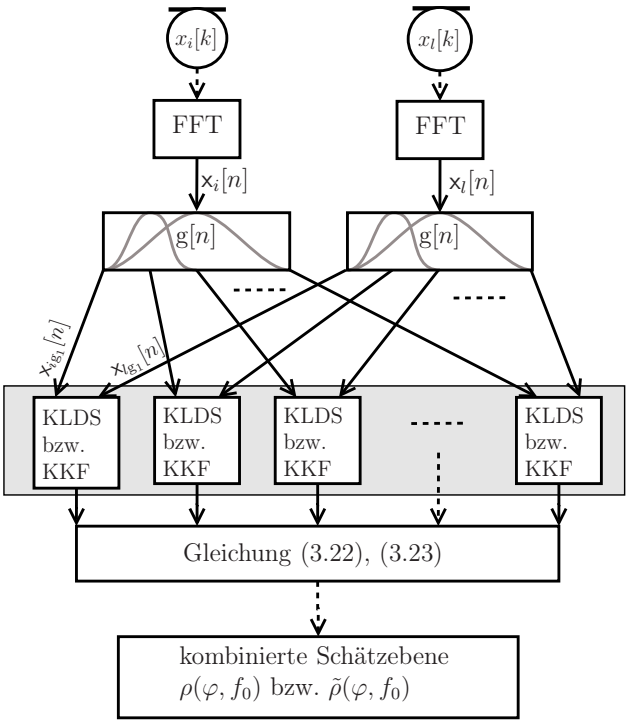
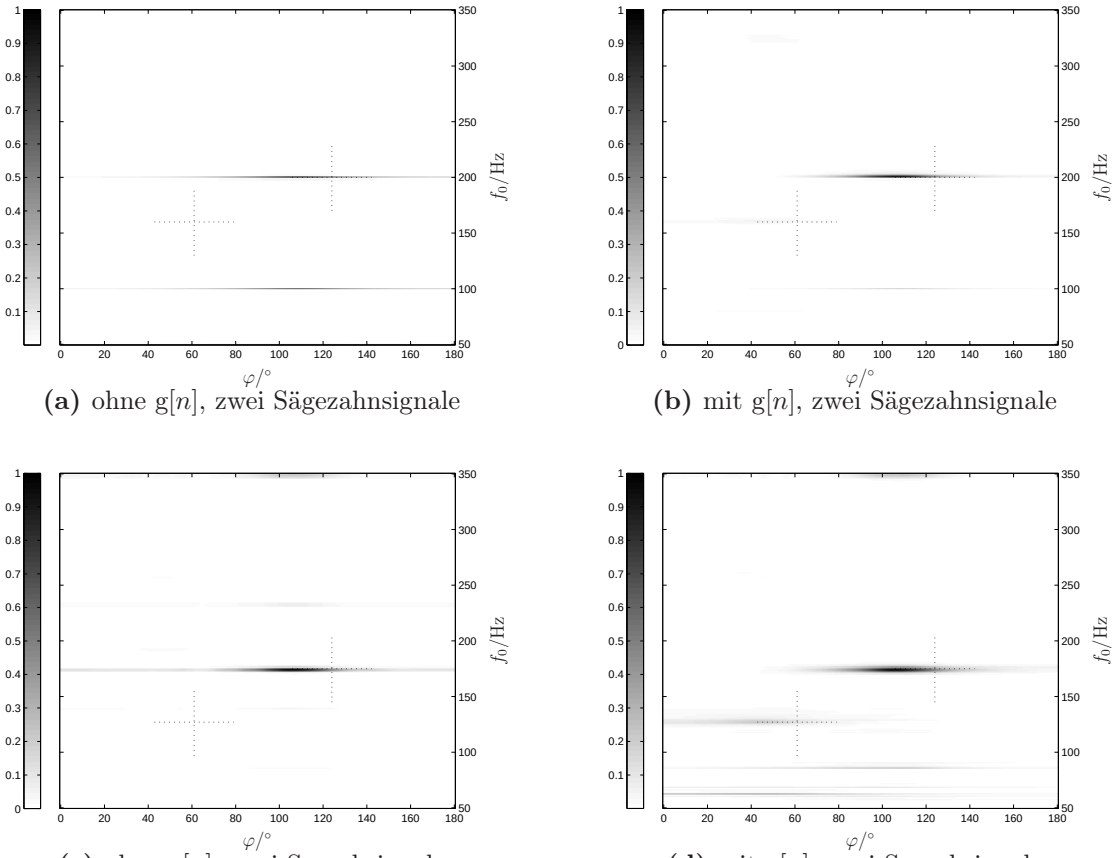


Abbildung 3.15: Flussdiagramm für die Bandpassfilterung anhand eines Mikrofonpaares. Aus dem Spektrum eines Mikrofonkanals werden F bandpassgefilterte Spektren erzeugt und mit dem entsprechenden Pendant des zweiten Mikrofons zu einer KKF oder LDS weiterverarbeitet, sowie anschließend wieder über Gleichung (3.22) oder (3.23) zu einer KKF bzw. LDS vereint.

zu trennen (vgl. Abschnitt 5.6). Dies bedeutet nicht, dass die Grundfrequenzen in unterschiedliche Bandpässe aufgetrennt werden, eher führen deren harmonische Vielfache in unterschiedlichen höheren Bändern zu einer deutlicheren Beeinflussung. Durch die anschließende Normierung der einzelnen Bänder und Aufsummierung kommt deren Einfluss in den höheren Bändern mehr zum Tragen [KOH08]. Da die Zielsetzung der Arbeit die Schätzung der Grundfrequenz auf menschliche Sprache begrenzt, kann davon ausgegangen werden, dass hohe Frequenzen ($> 8\,\mathrm{kHz}$) keine Information über die Sprachquellen enthalten. Die Begrenzung der Mittenfrequenzen auf maximal 8 kHz elementiert somit zusäzlich unerwünschte Hochtoneinflüsse aus der Schätzung.

Zur Veranschaulichung der Auswirkung der Filtervorverarbeitung sind in Abbildung 3.16 vier Ergebnisebenen aufgetragen. Die sich dahinter verbergenden Schätzverfahren unterscheiden sich nur in der Verwendung der Filtervorverarbeitung (rechte Seite). Als akustische Quellen dienten jeweils zwei ungestörte Sägezahnsignale (obere Diagramme) sowie zwei Sprachsignale (männlich und weiblich). Die Abbildungen auf der rechten Seite zeigen eine Verbesserung der Schätzungen gegenüber der linken Seite. Zudem scheint das Verfahren generell besser für reale Sprache als für deterministische Sägezahnsignale zu funktionieren. Anhand dieser Bilder lässt sich vermuten, dass die Grundfrequenzschätzung bei der kom-

binierten Auswertung robuster als die DoA-Schätzung zu sein scheint. Nähere Untersuchungen dazu sind in den Abschnitten 5.7 bis 5.9 zu finden.



(a) ohne g[n], zwei Sägezahnsignale **(b)** mit g[n], zwei Sägezahnsignale

(c) ohne g[n], zwei Sprachsignale **(d)** mit g[n], zwei Sprachsignale

Abbildung 3.16: Ergebnisebenen der kombinierten Schätzung. Jeweils zwei Quellen, deren wahre Charakteristika durch Kreuze gekennzeichnet sind. Auf der linken Seite ist die Schätzung ohne und auf der rechten Seite mit der Filterbankvorverarbeitung durchgeführt worden. Bei den Graphen **(a)** und **(b)** sind zwei Sägezahnsignale mit $f_0 = 160\,\text{Hz}$ bzw. $f_0 = 200\,\text{Hz}$ aus einer Richtung von 61° und 124° vorhanden. In den Graphen **(c)** und **(d)** handelt es sich um Sprachsignale aus den selben Richtungen mit einer weiblichen $f_0 = 175\,\text{Hz}$ und männlichen Stimme $f_0 = 127\,\text{Hz}$. Die Blocklänge betrug $N = 4096\,\text{Samples}$ bei einer Abtastrate von $f_s = 48\,\text{kHz}$. KLDS-T_1 Verfahren mit Cepstrum-Gewichtung, 15 Mikrofonpaare, MCCC-Kombination und 64 Bandpassgammatongewichtungen.

3.7 GCC-Phat Modifikationen

Die letzten Erweiterungen für das kombinierte Schätzverfahren führen dem Verfahren eine partielle Phat-Transformation (siehe Abschnitt 2.3.6) der KKF, beziehungsweise eine Phat-Gewichtung der resultierenden Ergebnisebene hinzu. Nach

[HOK08] ist die kombinierte Schätzung anfällig gegen Mehrwegeausbreitungen (vgl. Abschnitt 2.3.1) der Signalquellen. Um diesen Einfluss zu minimieren wird in [HOK08] vorgeschlagen eine partielle Phat-Transformation der KKF vorzunehmen. Durch die Phat-Transformation wird der Betrag des Leistungsspektrums zu eins gesetzt, wodurch die Korrelation ihre Periodizität verliert. Da jedoch die Grundfrequenz in genau dieser Periodizität codiert ist, muss für die kombinierte Grundfrequenz- und Richtungsschätzung eine Restriktion vorgenommen werden. Wie bereits in Abschnitt 2.3.5 erläutert, kann der für die Richtungsbestimmung relevante Teil der KKF auf den Bereich $\pm\kappa_{\max}$ (Gleichung (2.42)) beschränkt werden, sodass sich die neue KKF-Funktion zu

$$\hat{r}_{\text{hyb}}[\kappa] = \begin{cases} \hat{r}_{\text{phat}}[\kappa] & \text{, wenn } \kappa \in \langle -\kappa_{\max}, +\kappa_{\max} \rangle \\ \hat{r}[\kappa] & \text{, sonst} \end{cases} \tag{3.24}$$

ergibt. Um diese hybride Korrelation zu erzeugen, muss parallel eine gewöhnliche sowie Phat-transformierte Korrelation berechnet werden, aus den sich die hybride Variante $\hat{r}_{\text{hyb}}[\kappa]$ zusammensetzen lässt. Abbildung 3.17 illustriert das Verfahren in einem Flussdiagramm.

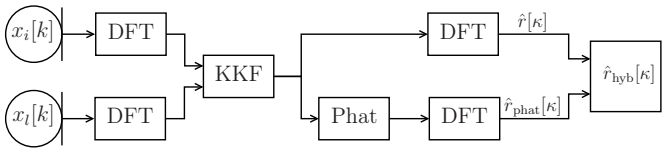
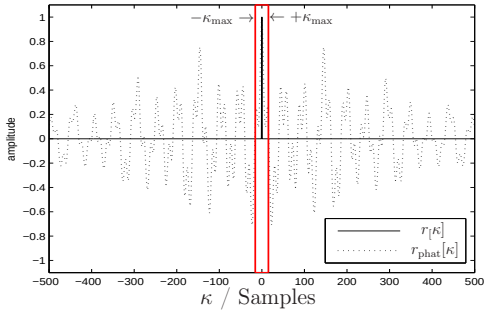
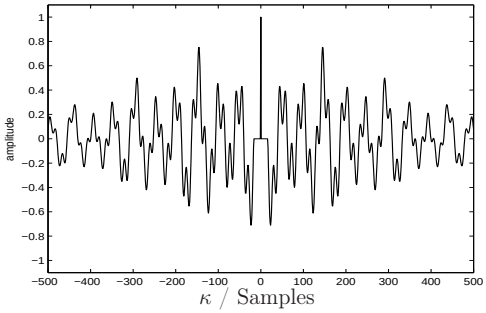


Abbildung 3.17: Flussdiagramm zur Erzeugung der hybriden KKF $\hat{r}_{\text{hyb}}[\kappa]$ aus zwei Mikrofonsignalen $x_i[k]$ und $x_l[k]$.

Somit bedeutet die hybride Korrelation einen Mehraufwand für die Berechnung, da zumindest Teile der Korrelation zweifach berechnet werden. Das Verfahren ist in Abbildung 3.18 für ein ungestörtes Sägezahnsignal aus frontaler Einfallsrichtung dargestellt.



(a) $\hat{r}_{\text{phat}}[\kappa]$ und $\hat{r}[\kappa]$ Sprachsignal



(b) $\hat{r}_{\text{hyb}}[\kappa]$ Sprachsignal

Abbildung 3.18: Kurzzeitkorrelationen 85 ms, **(a)** normale & phatgewichtete Korrelation, **(b)** kombinierte Variante, Einfallsrichtung $\varphi = 90°$. Ohne Nachhall oder Rauschen.

Das hybride Verfahren nach Gleichung (3.24) kann jedoch nicht für die kombinierte Schätzung über das KLDS angewandt werden. Bei der KLDS codiert sich die Einfallsrichtung der Signalquelle in der Phase, wodurch eine Beschränkung auf einen für die Richtungsschätzung relevanten Bereich nicht möglich ist. Daher wird im Folgenden eine eigne, neue und von der Cepstrum-Gewichtung inspirierte GCC-Phat-Gewichtung eingeführt. Bei dieser dient eine parallel zum KLDS berechnete GCC-Phat-Korrelation als Gewichtungsfunktion für das Ergebnisfeld der Schätzung:

$$\rho_{il,\mathrm{phat}}(\varphi, f_0) = \hat{r}_{x_i x_l,\mathrm{phat}}[\lfloor \kappa_{il}(\varphi) \rfloor] \cdot \rho_{il}(\varphi, f_0). \tag{3.25}$$

Aus der GCC-Phat-Funktion werden die Samples benötigt, deren Blockindizes in dem Bereich der betrachteten, möglichen Richtungsverzögerung liegen. Die Gewichtungsfunktion wird über alle Einfallswinkel, unabhängig von der Grundfrequenz, entsprechend der Lookup-Tabelle $\kappa_{il}(\varphi)$ auf das Ergebnisfeld pro Mikrofonpaar angewendet und kann daher unabhängig vom Schätzverfahren über das KLDS oder die KKF verwendet werden. Durch die nachträgliche Gewichtung werden Schätzungen hervorgehoben, die nur sowohl mit der Kernfunktion der Schätzung als auch durch die GCC-Phat-Funktion erkannt werden. Zur Verdeutlichung des Prinzips ist in Abbildung 3.19 in der oberen Grafik eine vollständige Phat gewichtete Korrelation eines Mikrofonpaares gezeigt, wohingegen die untere Abbildung die sich ergebende Gewichtung zeigt. An den Rändern der Gewichtungsfunktion ist zudem die winkelabhängige Genauigkeit der Richtungsschätzung per KKF (vgl. Abschnitt 3.2) erkennbar.

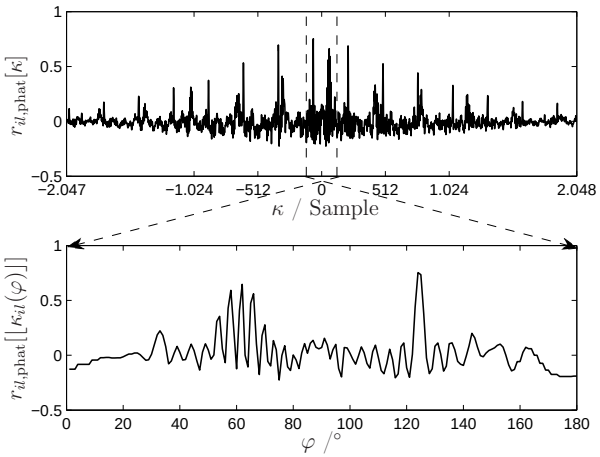
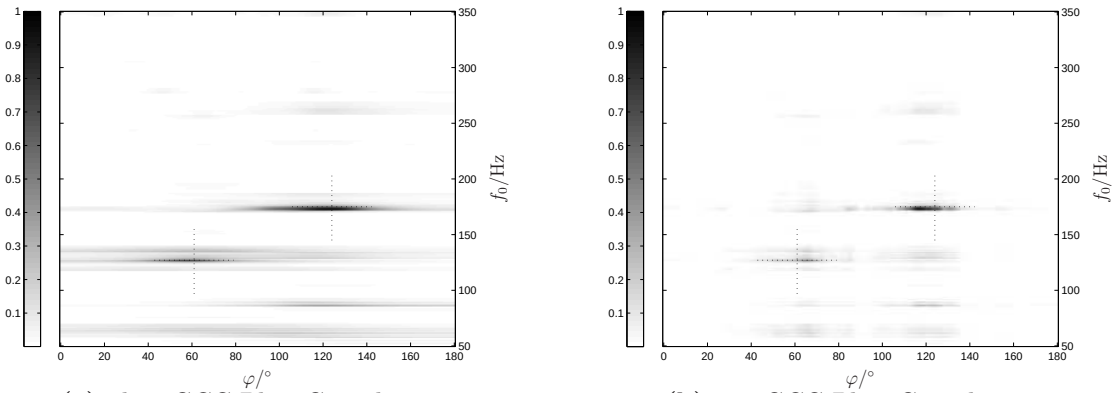


Abbildung 3.19: Obere Grafik: vollständige GCC-Phat-Funktion. Untere Grafik: aus der GCC-Phat über Gleichung (3.25) bestimmte Gewichtung für ein Mikrofonpaar mit dem Abstand $d_{il} = 0{,}88\,\mathrm{m}$ bei einer Abtastrate von $f_s = 48\,\mathrm{kHz}$. Die Sprecher befinden sich in Richtung $\varphi_0 = 61°$ und $\varphi_0 = 124°$.

Abbildung 3.20 zeigt abermals eine beispielhafte Kurzzeitschätzung einer akustischen Szene, diesmal jedoch mit implementierter GCC-Phat Modifikation bei einer simulierten Nachhallzeit von ca. 250 ms mit zwei simultanen Sprechern (männlich und weiblich). Die Version mit GCC-Phat Gewichtung (Abbildung 3.20b) zeigt eine deutlich verbesserte Richtungsschätzung. Obwohl beide Quellen eine identische Leistung besitzen, ist eine Schätzung dominanter ausgeprägt.



(a) ohne GCC-Phat Gewichtung **(b)** mit GCC-Phat Gewichtung

Abbildung 3.20: Zwei Sprachsignalquellen mit einer weiblichen $f_0 = 175\,\text{Hz}$ sowie männlichen Stimme $f_0 = 127\,\text{Hz}$ aus $\varphi_0 = 61°$ und $124°$, deren Charakteristika durch Kreuze gekennzeichnet sind. Auf der linken Seite ist die Schätzung ohne und auf der rechten Seite mit GCC-Phat-Gewichtung durchgeführt worden. Die Blocklänge betrug $N = 4096\,\text{Samples}$ bei einer Abtastrate von $f_s = 48\,\text{kHz}$. KLDS-T_1 Verfahren mit Cepstrum-Gewichtung, 15 Mikrofonpaare mit MCCC-Kombination sowie 64 Bandpassgammatongewichtungen.

Kapitel 4

Quellendetektion und -verfolgung

Das vorhergehende Kapitel beschreibt die in dieser Arbeit implementieren Verfahren für eine kombinierte Grundfrequenz- und Richtungsschätzung. Bis zu diesem Punkt wurde jedoch jeweils immer nur ein Kurzzeitausschnitt der Schätzung betrachtet. Außerdem wurde noch kein Bezug auf die automatische Auswertung der Ergebnisebenen genommen. Ein Problem stellt zudem die zeitliche Synthese der geschätzten Quellencharakeristiken von Block ℓ zu $\ell + 1$ dar. Um eben diese Probleme zu lösen, wird in diesem Kapitel das implementierte Partikel-Filter zur automatischen Quellenbestimmung und -verfolgung beschrieben. Vorteil dieser Methode gegenüber einer einfachen blockbasierten Maximumsuche und anschließender rekursiven Glättung der ermittelten Quellenchrakteristik ist die robustere Schätzung der Quelle(n) in akustisch komplexen Umgebungen. In solchen Situationen kommt es häufig vor, das Maxima der Schätzfunktion auf fehlerhafte Schätzungen durch Reflexionen zurückzuführen sind [WLW03]. Zudem enthält das Partikel-Filter bereits ein physikalisches Bewegungsmodell der zu schätzenden Quellen, wodurch die Verfolgung von sich bewegenden Quellen berücksichtigt wird.

Das Kapitel ist unterteilt in den Abschnitt 4.1 zur Erläuterung des eigentlichen Partikel-Filter-Prinzips. Abschnitt 4.2 befasst sich mit der in dieser Arbeit entwickelten „Sperr“-Filterung zur simultanen Mehrsprecherdetektion.

4.1 Partikel-Filterung

Die Schätzungen der kombinierten Einfallsrichtung und Grundfrequenz zeigen unter dem Einfluss von unerwünschten Störgrößen (z.B. Rauschen, Nachhall) einen verrauschten Charakter. Partikel-Filter, auch bekannt als *Sequenzielle Monte-Carlo-Methoden*, bieten trotz dieser verrauschten Messungen die Möglichkeit die Entwicklung eines dynamischen Systems, auf Grundlage des Bayestheorem, über einen Zeitraum zu schätzen [AMGC02]. Sie stellen eine praktikable Methode zum „Tracking“ (deut. Nachverfolgung) von einzelnen oder mehrfachen Sprechern dar und eignen sich besonders für Situationen mit Mehrwegeausbreitung sowie nichtlinearen und nicht-gaußschen Prozessen [LJ07, AMGC02]. Um Rückschlüsse auf das dynamische System schließen zu können, sind dabei zwei Modellebenen notwendig. Zum einen wird ein Modell zur Synthese des zeitlichen Verlaufs des untersuchten Systems benötigt (der Zustandsraum). Das dabei verwendete Zustands-

raummodell enthält Zustands-Vektoren (Partikel) $\mathbf{s}$ die alle notwendigen Informationen zur Beschreibung des beobachteten Systems enthalten. Zum anderen wird eine Repräsentation des aktuellen Systemzustands aus aktuellen Messdaten ρ (hier: kombinierte Ergebnisebene) benötigt. Die zu den Partikeln gehörenden Gewichte w stellen die Verbindung des Zustandraumes mit den Messwerten dar.

Das Partikel-Filter unterteilt sich im Groben in zwei wesentliche Verarbeitungsschritte, der Vorhersage und der Aktualisierung. Die Aufgabe des Vorhersageschrittes ist es, die Partikelzustände sequentiell von Verarbeitunsblock ℓ zu $\ell + 1$ auf Grundlage eines physikalischen Bewegungsmodells vorherzusagen. Dieser Zustand der Partikel wird auch als A-priori-Verteilung bezeichnet. Während der anschließenden Aktualisierung werden die Gewichte w der Partikel an die realen Bedingungen über die Likelihood-Funktion $\mathrm{F}(\rho, \mathbf{s})$ (hervorgehend aus den realen Messungen) angepasst und somit die A-posteriori-Verteilung der Partikel erzeugt. Die Likelihood-Funktion dient in diesem Zusammenhang als bedingte Auftrittswahrscheinlichkeit $p(\rho|\mathbf{s})$.

Die Zustands-Raum-Beschreibung einer akustischen Umgebung bei der Partikel-Filterung beruht auf der Annahme, dass sich Schätzungen in der Ergebnisebene, hervorgerufen durch Signalquellen, einem dynamischen Modell entsprechend fortbewegen. Wohingegen Ausreißer durch Reflexionen oder andere Störungen keine zeitliche Konsistenz aufweisen [HR10]. Im Folgenden wird der Ablauf eines Partikel-Filters für die kombinierte Grundfrequenz und DoA-Schätzung erläutert.

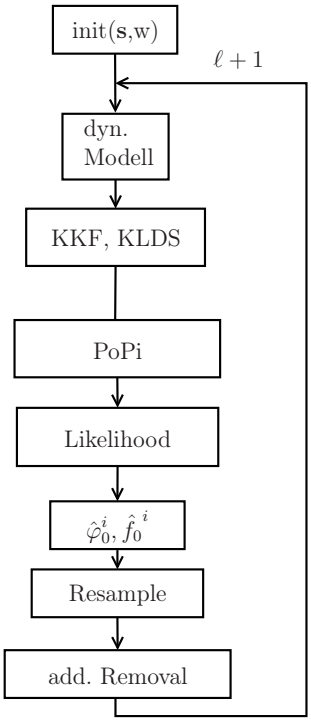


Abbildung 4.1: Flussdiagramm der Partikel-Filterung.

Abbildung 4.1 illustriert zur Übersicht die Partikel-Filterung für die kombinerte Grundfrequenz- und Richtungsschätzung in einem Flussdiagramm, beginnend mit

der Initialisierung der Partikel

$$\mathbf{s}_p^i = [x, y, z, \dot{x}, \dot{y}, \dot{z}, f_0] \tag{4.1}$$

und deren Gewichte

$$\mathrm{w}_p^i = \frac{1}{L} \qquad p = 1 \ldots L. \tag{4.2}$$

Die Partikel $\mathbf{s}_p^i$ definieren jeweils einen Zustand über die Quellenposition $[x, y, z]$ in kartesischen Koordinaten, der Quellengeschwindigkeit $[\dot{x}, \dot{y}, \dot{z}]$ sowie der Grundfrequenz f_0. Es werden jeweils $p = 1 \ldots L$ Partikel für $i = 1 \ldots \mathsf{Q}$ Quellen erzeugt. Die zu den Partikeln gehördenden Gewichte w_p^i werden zu Beginn mit $\frac{1}{L}$ gleichverteilt.

Im nächsten Schritt werden die Zustandsänderungen der Partikel von einem Zeitpunkt zum Nachfolgenden mit einem dynamischen Modell prognostiziert. In dieser Arbeit wurde als physikalisches Bewegungsmodell das *Langevin-Modell* umgesetzt [WLW03]. Dabei werden die Positionsänderungen in allen Dimensionen als unabhängige Prozesse erster Ordnung angesehen. Die Vorhersage für die Bewegung in der x-Achse sowie der Grundfrequenz (nicht Langevin-Modell) gestaltet sich wie folgt:

$$\dot{x}_\ell = a_x \dot{x}_{\ell-1} + b_x u, \tag{4.3}$$

$$x_\ell = x_{\ell-1} + \Delta T \dot{x}_\ell \qquad \Delta T = \frac{N}{f_s}, \tag{4.4}$$

$$a_x = \exp^{\beta_x \Delta T}, \tag{4.5}$$

$$b_x = v_x \sqrt{1 - a_x^2}, \tag{4.6}$$

$$f_{0,t} = f_{0,t-1} + b_f \cdot u, \qquad b_f = \Delta T \cdot \beta_{f_0}. \tag{4.7}$$

Die Variable u ist dabei eine normalverteilte Zufallsgröße und ΔT spiegelt die Zeitperiode zwischen zwei Vorhersagen wieder. Für die Vorhersage der Grundfrequenz wurde keine begrenzende Beschleunigung eingeführt, da davon auszugehen ist, dass sich diese instantan ändern kann.

Anschließend wird aus der KKF bzw. KLDS die kombinierte Ergebnisebene $\rho'(\varphi, f_0)$ bzw. $\tilde{\rho}'(\varphi, f_0)$ gemäß Kapitel 3 berechnet. Das Ergebnis dient der Partikel-Filterung als Pseudo-Likelihood-Funktion $\mathrm{F}(\rho'^i, \mathbf{s}_p^i)$,

$$\mathrm{F}(\rho'^i, \mathbf{s}_p^i) = \max(\rho^i(\mathbf{s}_p^i), 0)^{\mathrm{r}} \qquad i = 1 \ldots \mathsf{Q}, \tag{4.8}$$

wobie diese Funktion für jede zu ermittelnde Quelle i getrennt zu erstellen ist. Mit der Variable $\mathrm{r} \in \mathbb{R}$ ist noch eine Beeinflussung von $\mathrm{F}(\rho^i, \mathbf{s}_p^i)$ möglich. In dieser Arbeit ist $\mathrm{r} = 2$ gesetzt worden, um eine größere Betonung auf hohe Wahrscheinlichkeiten zu legen. Bevor Gleichung (4.8) mit den einzelnen Partikeln berechnet werden kann, muss die Quellenposition aus den kartesischen Koordinaten $[x, y, z]$

über den euklidischen Abstand zum Bezugspunkt des Mikrofonarrays in einen Einfallswinkel $\hat{\varphi}_0$ umgerechnet werden. Bei einem Mikrofonpaar berechnet sich der Einfallswinkel wie folgt:

$$\hat{\varphi}_{0,\mathbf{s}} = \arccos\left(\frac{\|l_\mathbf{s} - l_1\|_2 - \|l_\mathbf{s} - l_2\|_2}{\|l_1 - l_2\|_2}\right). \tag{4.9}$$

Zur Veranschaulichung von Gleichung (4.9) wird das darin enthaltene Vorgehen in Abbildung 4.2 nocheinmal dargestellt. Jedes Partikel $\mathbf{s}_p$ wird so zunächst als potentielle Quellenausprägung betrachtet.

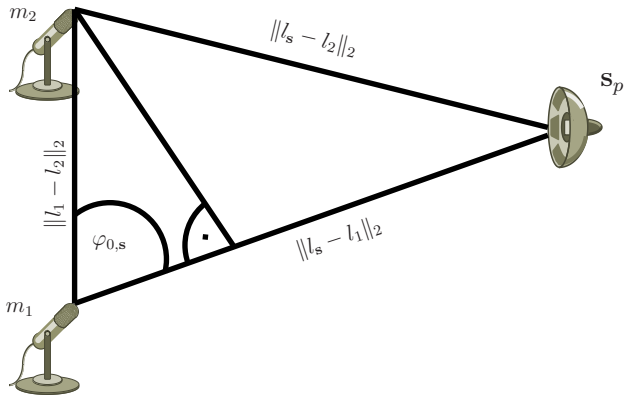


Abbildung 4.2: Bestimmung des Einfallswinkels eines Partikel ausgehend von kartesischen Koordinaten.

Die Variable $l_\mathbf{s} = [x, y, z]$ umfasst die Koordinaten des Partikels, l_1 und l_2 beinhalten entsprechend die Koordinaten der Mikrofone. Die Likelihood-Funktion dient im Folgenden als bedingte Auftrittswahrscheinlichkeit,

$$p(\rho'^{,i}|\mathbf{s}_p^i) = \mathrm{F}(\rho'^{,i}, \mathbf{s}_p^i), \tag{4.10}$$

über die sich die Gewichtungen

$$\mathrm{w}_p^i = p(\rho'^{,i}|\mathbf{s}_p^i) \tag{4.11}$$

der einzelnen Partikel bestimmen lassen, welche nach einer Normierung pro Signalquelle in der Summe 1 ergeben.

$$\sum_{p=1}^{L} \mathrm{w}_p^i = 1 \tag{4.12}$$

Aus den normierten Gewichtungen sowie den einzelnen Partikeln kann die Schätzung für Quelle i ermittelt werden,

$$(\hat{\varphi}_0^i, \hat{f}_0^{\,i}) = \sum_{p=1}^{L} \mathrm{w}_p^i \mathbf{s}_p^i. \tag{4.13}$$

Die Änderung des Systemzustandes ist von Verarbeitungsblock ℓ zu ℓ von Störungen überlagert, welche durch Rauschen bei der Vorhersage modelliert werden (Bestandteil des physikalischen Bewegungsmodells). Daher verursacht die Vorhersage eine Aufweitung der Wahrscheinlichkeitsdichtefunktion (Partikelzustände). Das Auseinanderdriften der Partikel wird auch als Degeneration bezeichnet [AMGC02]. Zur Vermeidung der Degeneration des Partikel-Filters wird ein sogenanntes Resampling durchgeführt. Das hier verwendete Verfahren des *Systematischen Resamplings* zur Unterdrückung diese Phänomens ist in [AMGC02] näher erläutert. Dabei werden im Verlauf des Resamplings Partikel mit hoher Gewichtung vervielfältigt, wohingegen Partikel mit geringer Gewichtung beseitigt werden. In Hinblick auf eine effektive Berechnung wird dieser Abschnitt jedoch nur bei Unterschreitung eines Schwellwertes L_T berechnet.

$$\hat{L}_{eff} < L_T, \tag{4.14}$$

$$\hat{L}_{eff} = \frac{1}{\sum_{p=1}^{L} (\mathrm{w}_p^i)^2}. \tag{4.15}$$

Der Entscheidungswert $\hat{L}_{eff}$ ist eine Schätzung der effektiven Stichprobengröße L_{eff} [AMGC02]. Der Schwellwert wurde auf $L_T = \frac{L}{10}$ gesetzt. So dass, wenn die Summe der quadrierten Gewichte (Gleichung (4.15)) kleiner als 10% der Partikelanzahl ist, ein Resampling durchgeführt wird. Die Summe der Gewichte deutet zu diesem Zeitpunkt auf eine ungenügende Fokussierung des Partikel-Filters auf hohe Quellenschätzungen hin, sondern eher auf eine hohe Verschmierung der Partikel-Zustände.

Nach erfolgter Bestimmung aller Quellencharakteristiken wurde ein zusätzlicher Verarbeitungsschritt implementiert (vgl. Abbildung 4.1 „add. removal“), wie er auch in [HR10] beschrieben wird. Dabei wird ein weiterer Teil der Partikel jeder Quelle gelöscht und zufällig in der Nähe von nicht mehr als 5 lokalen Maxima der unbearbeiteten Ergebnisebene neu positioniert. Die Maxima werden nach einer zweidimensionalen Glättung aus der Ergebnisebene $\rho(\varphi, f_0)$ ermittelt. So kann ausgeschlossen werden, dass nicht zufällig dicht beieinander liegende Maxima als Anhaltspunkt für die neuen Zustände herangezogen werden. Der Abstand zwischen zwei Maxima wurde hier mit mindestens $f_0 = 10\,\mathrm{Hz}$ und $\varphi = 10°$ festgelegt. In Abbildung 4.3 wird das Verfahren beispielhaft illustriert. Die gefundenen Maxima müssen nicht zwangsläufig eine wahre Quelle repräsentieren. Die Menge der zu löschenden Partikel wurde nach Vorlage von [HR10] auf 20 % gesetzt. Ziel des Verarbeitungsschrittes ist es, den Algorithmus robuster für die Mehrsprechererkennung zu machen, indem ein Teil der Partikel auch auf kleinere Maxima verschoben wird. Durch die Sperr-Filterung ist nicht immer sichergestellt, dass die gesuchte Quelle ein dominantes Maximum in der Ergebnisebene erzeugt. So kann die Schätzung der konkurierenden Quellen fehlerhaft sein und das Sperr-Filter die zugehörigen Maxima nicht auslöschen. Oder durch Reflexionen verursachte Maxima besitzen eine höhere Ausprägung als die wahre Quelle.

Nach erfolgter Berechnung aller Quellenschätzungen für den aktuellen Zeitpunkt ℓ und den anschließenden Nachverarbeitungen durch Resampling und zusätzliches

Verschieben der Partikel springt das Verfahren wieder zurück zur Vorhersage der Partikelbewegungen und wird mit Zeitschritt $\ell + 1$ fortgeführt.

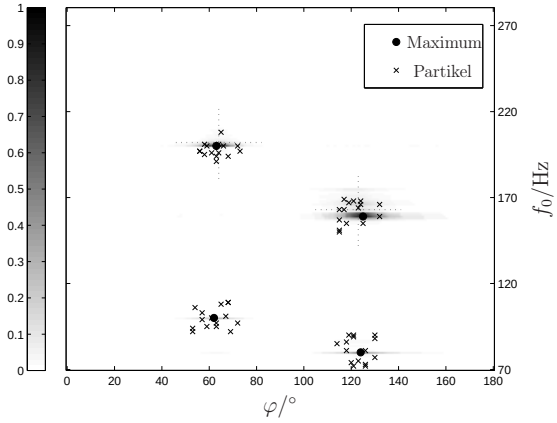
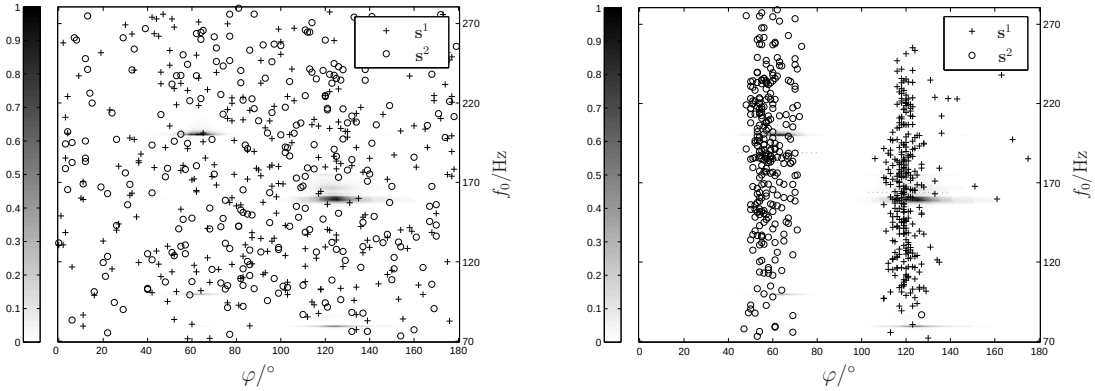


Abbildung 4.3: Ergebnisebene mit Maximumsuche, sowie Partikelneuanordnung („add. Removal“). Die wahren Signalquellen sind jeweils ungestörte Sägezahnsignale aus $\varphi_0 = 64°, 121°$ mit Grundfrequenzen von $f_0 = 200\,\text{Hz}, 160\,\text{Hz}$. Die Kreuze zeigen die durch das Partikel-Filter ermittelten Signalquellen nach sechs Iterationen, $f_s = 48\,\text{kHz}$, $N = 4096\,\text{Samples}$.



(a) Partikelverteilung nach der Initialisierung.

(b) Partikelverteilung nach sieben Iterationen.

Abbildung 4.4: Ergebnisebenen mit überlagerter Darstellung der Partikel für zwei Quellen. Anzahl der Partikel $L = 300$. **(a)** Inizialisierungszustand der Partikel **(b)** Partikelverteilung nach sieben Iterationen für die vorliegende akustische Szene mit zwei simultanen Sägezahnsignalen aus $\varphi = 64°, 121°$ mit Grundfrequenzen von $f_0 = 200\,\text{Hz}, 160\,\text{Hz}$. $f_s = 48\,\text{Hz}$ bei $N = 4096$. Partikelvorhersage für die Grundfrequenzschätzung $\beta_{f_0} = 50\,\text{Hz}$ ist augenscheinlich etwas zu groß gewählt.

Abbildung 4.4 illustriert für ein Beispiel mit zwei statischen simultanen Sägezahnsignalquellen $f_{0,1} = 200\,\text{Hz}$ und $f_{0,2} = 160\,\text{Hz}$ aus Richtung $\varphi_{0,1} = 64°$ und

$\varphi_{0,2} = 121°$ die Verteilung der Partikel unmittelbar nach der Initialisierung (linke Seite) sowie die Verteilung nach der siebten Iteration.

4.2 Sperr-Filterung

Aufgabe der Sperr-Filterung im Verlauf der Quellenerkennung ist es den Einfluss nur einer Quelle in der KKF bzw. KLDS zu maximieren. Simultane Sprecher in der Ergebnisebene der kombinierten Schätzung beinflussen sich gegenseitig, in dem eine dominate Quelle weitere Quellen in der Ergebnisebene unterdrückt und diese dann nicht als eigene Quelle detektiert werden kann. Dazu wird versucht die durch simultane Quellen verursachten zusätzlichen Einflüsse zu eleminieren.

Abbildung 4.5 zeigt das Flussdiagramm des Partikel-Filters erweitert um die Sperr-Filterung. Im dritten Arbeitsschritt des Flussdiagramms werden die Mikrofonsignale, bedingt durch die verwendete Kernfunktion zur Grundfrequenz- und Richtungsschätzung, in eine KKF bzw. KLDS umgewandelt. Auf diese Daten wird anschließend die „Sperr"-Filterung angewandt. In Abhängigkeit von der Anzahl verfolgter Quellen muss die Filterung mehrmals durchgeführt werden.

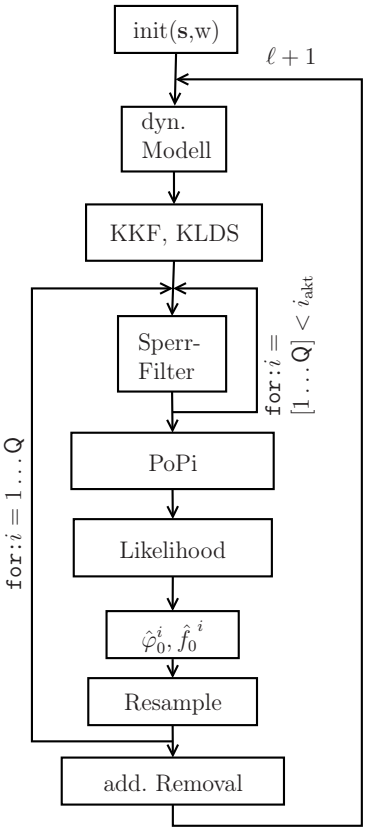


Abbildung 4.5: Erweitertes Flussdiagramm der Partikel-Filterung mit Sperr-Filterung. Q definiert die Anzahl der Signalquellen.

Es wurden zwei verschiedene Sperr-Filter für den Zeitbereich sowie für den Fre-

quenzbereich (Bandstopp-Filter) implementiert. Das Verfahren für den Zeitbereich [Vol10, BOS08] manipuliert die KKF mit:

$$\breve{r}_{x_i x_l}[\kappa] = \hat{r}_{x_i x_l}[\kappa] \cdot \Psi_{il}[\kappa], \tag{4.16}$$

$$\Psi_{il}[\kappa] = \sum_{p=-P}^{P} 1 - \exp \frac{|\kappa - \lfloor p \cdot \nu_0(\hat{f}_0) + \kappa_{il}(\hat{\varphi}_0) \rfloor|}{br_{il}(\beta_{br})}, \tag{4.17}$$

$$br_{il}(\beta_{br}) = -\frac{d_{il} \sin(\beta_{br}) f_s}{c\, 2 \ln(0.5)}. \tag{4.18}$$

Wobei die Sperr-Funktion $\Psi_{il}[\kappa]$ aus den geschätzten Quellenparametern $\hat{f}_0$ und $\hat{\varphi}_0$ für jedes Mikrofonpaar getrennt erzeugt wird. Der Parameter $br_{il}(\beta_{br})$ bestimmt die Breite des einzelnen Kerben für den Bereich in Abhängigkeit vom Mikrofonabstand d_{il} der Abtastfrequenz f_s und der Schallgeschwindigkeit c und lässt sich über β_{br} variieren[1] [Vol10, BOS08].

Das Verfahren für den Frequenzbereich manipuliert den Betrag des KLDS mit

$$|\breve{\Phi}_{x_i x_l}[n]| = |\hat{\Phi}_{x_i x_l}[n]| \cdot \boldsymbol{\Psi}[n], \tag{4.19}$$

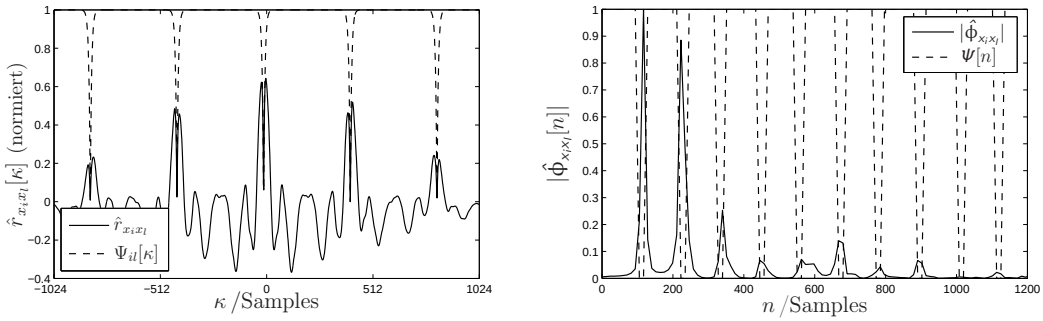
$$\boldsymbol{\Psi}[n] = \begin{cases} 0 & \text{, bei } p \cdot n(\hat{f}_0) \pm \beta_{br} \qquad \text{, mit } p = 1 \ldots P \\ 1 & \text{, sonst} \end{cases}. \tag{4.20}$$

Im Unterschied zur Zeitbereichsfilterung fließt hier nur die Schätzung der Grundfrequenz $\hat{f}_0$ in das Sperr-Filter ein. Dies hat den Vorteil, dass die Filterfunktion $\boldsymbol{\Psi}[n]$ nur einmal pro Zeitabschnitt und zu unterdrückender Quelle zu berechnen ist und anschließend für alle Mikrofonpaare geeignet ist. Durch Unterdrückung des Betrages der KLDS hat die Phase für diese Frequenzen in der kombinierten Schätzung keinen Einfluss mehr. Zwei simultane Quellen mit der gleichen Grundfrequenz aber unterschiedlicher DoA sind jedoch bei der Berechnung mittels KLDS prinzipbedingt nicht möglich. Die Variable P definiert jeweils die Anzahl zu erzeugender Kerben. Als Vorlage für die Sperr-Filter dienen die Schätzungen der konkurierenden Quellen im selben Iterationszeitpunkt ℓ. Ist diese noch nicht vorhanden, wird die Schätzung aus dem Zeitpunkt $\ell - 1$ verwendet.

Anhand der manipulierten KKF bzw. KLDS kann die Eregebnisebene für die kombinierte Schätzung neu berechnet werden. Diese dient dann wiederum als Pseudo-Likelihood-Funktion für das Partikel-Filter einer weiteren zu schätzenden Quelle. Um eine weitere simultane Signalquelle zu detektieren springt der Partikel-Filter Algorithmus zurück zum Verarbeitungsschritt der Sperr-Filterung. In dieser Iteration werden jedoch andere quellenbezogene Auslöschungen vorgenommen, sodass eine Ergebnisebene der kombinierten Schätzung erzeugt wird, in der die Quelle $i + 1$ im Idealfall dominat hervorgehoben wird. Zur Bestimmung der Quelle $i + 1$ wird wie auch für Quelle i jeweils eine neue Liklihood-Funktion erzeugt und mit den der Quelle $i + 1$ zugehörigen Partikel berechnet.

So bestimmte Ergebnisebenen sind in Abbildung 4.7 dargestellt. Die obere Abbildung 4.7a zeigt die Ergebnisebene vor der Sperr-Filterung. Die Kreuze zeigen

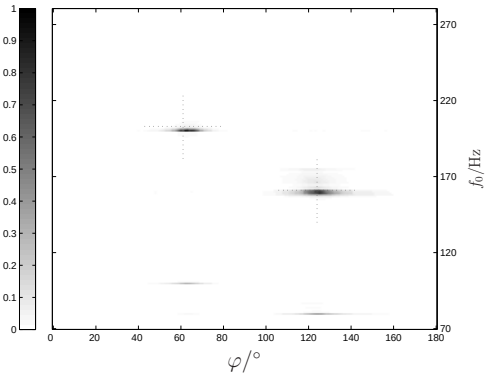
[1] Wenn nicht anders angebeben, ist in dieser Arbeit $\beta_{br} = 10$ gesetzt.



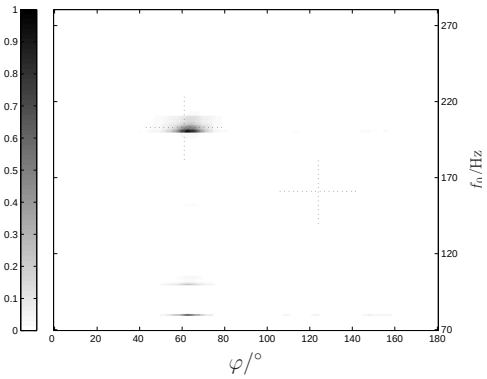
(a) Sperr-Filterung $\Psi_{il}[\kappa]$ anhand einer KKF. **(b)** Sperr-Filterung $\Psi[n]$ anhand eines KLDS.

Abbildung 4.6: Darstellung der Sperr-Filterung für den Zeitbereich **(a)** sowie Freyquenzbereich **(b)**. Als Vorlage diente eine Sägezahnquelle ($\varphi_0 = 64°$, $f_0 = 150\,\mathrm{Hz}$) aufgenommen von einem Mikrofonpaar

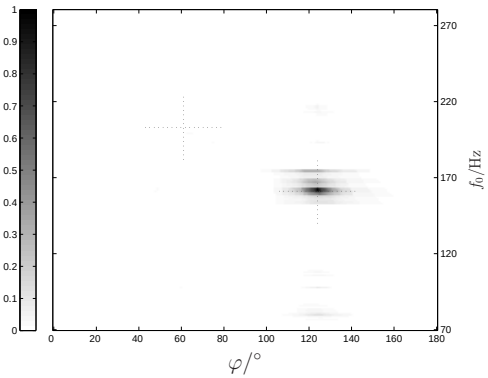
die durch vorherige Iterationen bestimmte Quellenschätzungen. Nach optimaler Anwendung der Sperr-Filterung auf die Ausgangs-KKF bzw. KLDS für jeweils eine Quelle ergeben sich die Abbildungen 4.7b und 4.7c. Der auslöschende Effekt erstreckt sich sowohl auf die wahre Grundfrequenz als auch auf deren harmonische Verwechslungen.



(a) ohne Sperr-Filterung



(b) Sperr-Filterung der Quelle 1



(c) Sperr-Filterung der Quelle 2

Abbildung 4.7: Ergebnisebenen $\tilde{\rho}'_{il}(\varphi, f_0)$ für zwei ungestörte Sägezahnsignale aus $\varphi_0 = 64°, 121°$ mit Grundfrequenzen von $f_0 = 200\,\text{Hz}, 160\,\text{Hz}$. Die gestrichelten Kreuze zeigen die durch das Partikel-Filter ermittelten Signalquellen nach sechs Iterationen. **(a)** Unbearbeitetes Ergebnisfeld. **(b)** Sperr-Filterung der Quelle 1 ($\varphi_0 = 121°$, $f_0 = 160\,\text{Hz}$), **(c)** Sperr-Filterung der Quelle 2 ($\varphi_0 = 64°$, $f_0 = 200\,\text{Hz}$). $f_s = 48\,\text{kHz}$, $N = 4096\,\text{Samples}$.

Kapitel 5

Evaluation

In diesem Kapitel werden die zuvor beschriebenen Algorithmen auf ihre Leistungsfähigkeit untersucht. Im ersten Abschnitt 5.1 wird der verwendete Mess- und Simulationsaufbau vorgestellt. Anschließend beschreibt Abschnitt 5.2 das *Acc*-Bewertungskriterium zur Feststellung der erzielten Trefferraten. Bezugnehmend auf die *Acc*-Ergebnisse werden in den Abschnitten 5.3 bis 5.6 die vorgestellten Algorithmen untersucht und jeweils die effektivste Einstellung ermittelt. Als Besonderheit, verglichen mit den in der Literatur gezeigten Evaluationen, wird hier neben der Richtungsschätzung immer auch die Grundfrequenzschätzung zur Auswertung herangezogen. In den letzten Abschnitten 5.7 und 5.8 wird die favorisierte Kombination aller Verfahren in unterschiedlichen akustischen Bedingungen, sowie erstmals für sich bewegende Quellen, ausgewertet. Weiterhin wird das eigene, optimierte Verfahren mit dem wohlbekannten GCC-Phat Verfahren als Pseudo-Likelihood-Funktion für das Partikel-Filter zur Richtungsschätzung verglichen.

5.1 Mess- und Simulationsaufbau

Der zur Evaluation der Algorithmen verwendete Testaufbau (Abbildung **??**) besteht aus einem Mikrofon-Line-Array mit insgesamt sechs Mikrofonen (Gleichung (5.1)) in einem Konferenzraum wie in Abbildung 5.1 dargestellt. Mit diesen sechs Kanälen können maximal 15 Mikrofonpaare gebildet werden. Die zu detektierenden Sprecher wurden nacheinander durch einen Lautsprecher an neun verschiedenen Positionen gemäß (5.3) repräsentiert,

$$l_{Mic\,m} = [0{,}33\,\mathrm{m}\ ,\ 2{,}20\,\mathrm{m} + 0{,}22\,\mathrm{m} \cdot (m-1)\ ,\ 1{,}12\,\mathrm{m}], \tag{5.1}$$

$$m = 1 \ldots 6, \tag{5.2}$$

$$l_{Quelle\,i} = [3{,}6\,\mathrm{m}\ ,\ 0{,}5\,\mathrm{m} \cdot i\ ,\ 1{,}12\,\mathrm{m}], \tag{5.3}$$

$$i = 1 \ldots 9. \tag{5.4}$$

Auf Grundlage dieses Aufbaus wurden reale Impulsantworten des Konferenzraumes aufgenommen und eine Nachhallzeit von $\tau_{60} \approx 550\,\mathrm{ms}$ ermittelt. Alle bei den Raumimpulsantwortmessungen verwendeten Geräte sind in Tabelle 5.1 aufgelistet und Abbildung 5.1b zeigt ihre schematische Verschaltung. Als Messverfahren

für die Raumimpulsantworten wurde eine Sinussweep-Verfahren [MM01] mit folgenden Parametern verwendet:

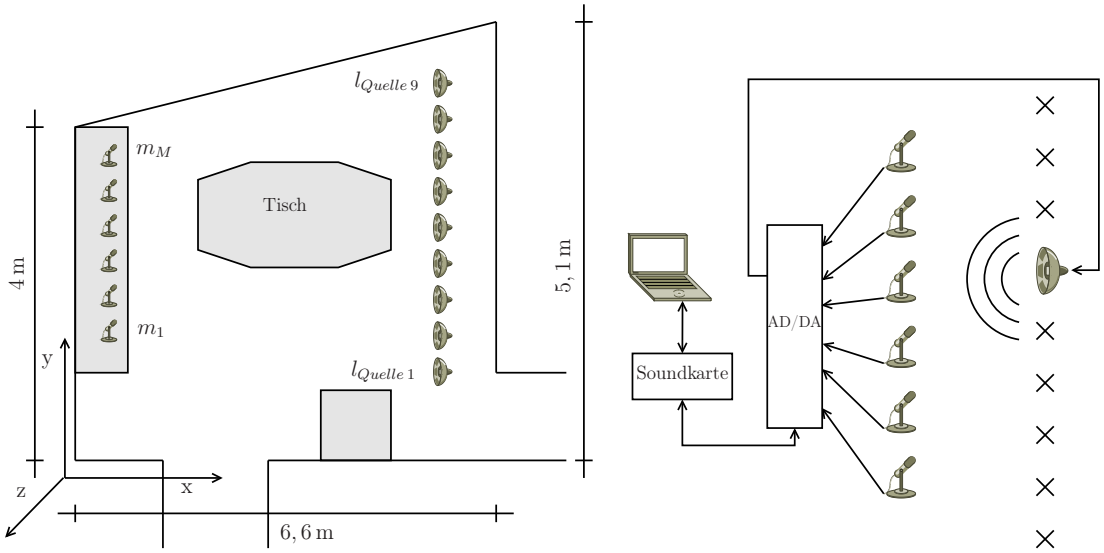
- Abtastfrequenz $f_s = 48\,\text{kHz}$
- 24 Bit Auflösung
- Länge von fünf Sekunden
- 25 Wiederholungen pro Quellenposition

Tabelle 5.1: Verwendete Hardware

Gerät	Typ
Mikrofon	6x Thomann SC180
AD/DA Wandler	Behringer ADA8000 PRO-8 DIGITAL
Soundkarte	RME HDSPE RAYDAT
Lautsprecher	1x Genelec 8020B
Computer	Dell Optiplex 755, Intel Quadcore 2,4 GHz

Zusätzlich wurde die Konstellation nach Gleichung (5.1) und (5.3) in einem Spiegelquellenmodell [Hab10] zur Simulation der Raumimpulsantworten mit unterschiedlichen Nachhallzeiten rekonstruiert. Als Raumgeometrie wurde ein rechteckiger Raum mit den Abmessungen [4,6 m; 5,1 m; 2 ,5 m] angenähert an den realen Raum verwendet und durch Variation der Reflexionskoeffizienten die Nachhallzeit verändert:

$$\tau_{60} = \{0, 20, 40, 70, 100, 250, 500\} \quad \text{in ms.} \tag{5.5}$$



(a) Physikalischer Messaufbau **(b)** Verbindung der Messgeräte

Abbildung 5.1: Darstellung des zur Evaluation verwendeten Messaufbaus. Konferenzraum der Fraunhofer Projektgruppe HSA im Haus des Hörens/ Oldenburg.

In Abbildung 5.2 sind eine simulierte und gemessene Impulsantwort für ähnliche Nachhallbedingungen abgebildet. Das Bild verdeutlicht, dass die mit simulierten RIR erzeugten Ergebnisse nicht unmittelbar mit Ergebnissen aus gemessenen RIR verglichen werden können.

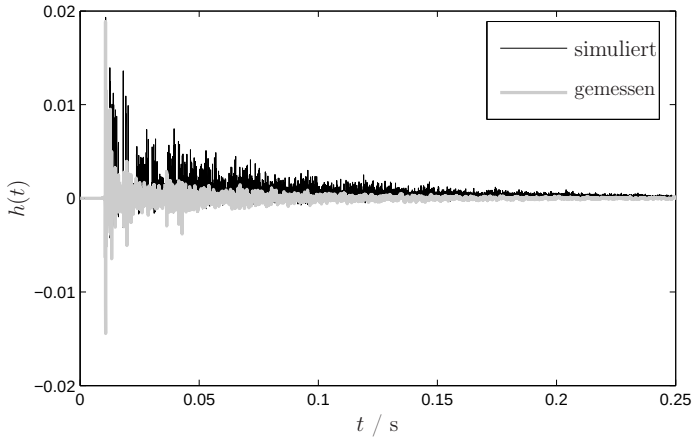


Abbildung 5.2: Vergleich von simulierter (schwarz) und gemessener (grau) RIR für ähnliche Bedingungen.

Für die Testszenarien mit Störgeräusch wurde ein sprachgefärbtes Rauschen auf Grundlage einer weiblichen Stimme verwendet. Um ein möglichst diffuses Störgeräuschfeld zu erzeugen, dienten alle neun Signalquellen stets simultan als Rauschquellen. Dabei wurde das Rauschen unabhängig von der für die Nutzsignale verwendeten RIR immer mit den realen, aufgenommenen RIR gefaltet. Abbildung 5.3 illustriert die sich für das erzeugte Rauschfeld einstellende Magnitude Squared Coherence (deut. Betragsquadrat der Kohärenzfunktion) (MSC) für ein Mikrofonpaar im Vergleich zur theoretisch berechneten MSC-Kurve für ein ideal diffuses Geräuschfeld [Bit01, GKM06, VHH98].

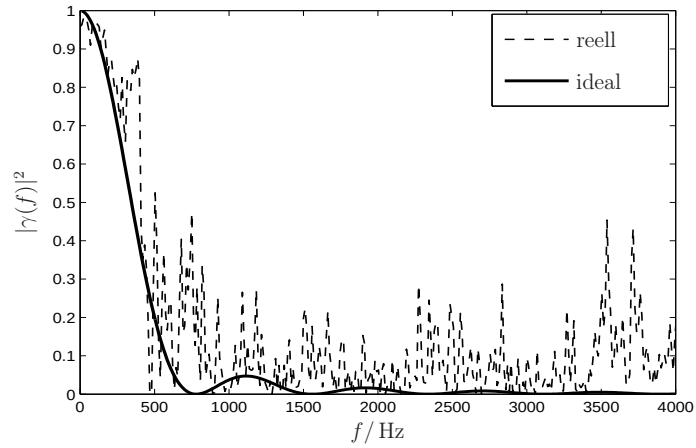


Abbildung 5.3: MSC des generierten Störgeräuschfeldes für einen Mikrofonabstand von $d = 0,22\,\mathrm{m}$.

Die Auswahl der in den nachfolgenden Abschnitten gewählten Testszenarien geschah jeweils unter dem Aspekt der besten Differenzierbarkeit des zu erörternden Sachverhalts. Nichts desto trotz wurde versucht eine Konstellation zu wählen, die möglichst realen Bedingungen nahe kommt.

5.2 Bewertungskriterium

Zum Vergleich der unterschiedlichen Erweiterungen und Algorithmen wurde ein Bewertungskriterium verwendet, was eine Aussage über die Trefferrate der Algorithmen zulässt [HKO08]:

$$Acc_{f_0} = \frac{1}{P}\sum_{\ell=1}^{P}\delta(\hat{f}_0, f_0) \tag{5.6}$$

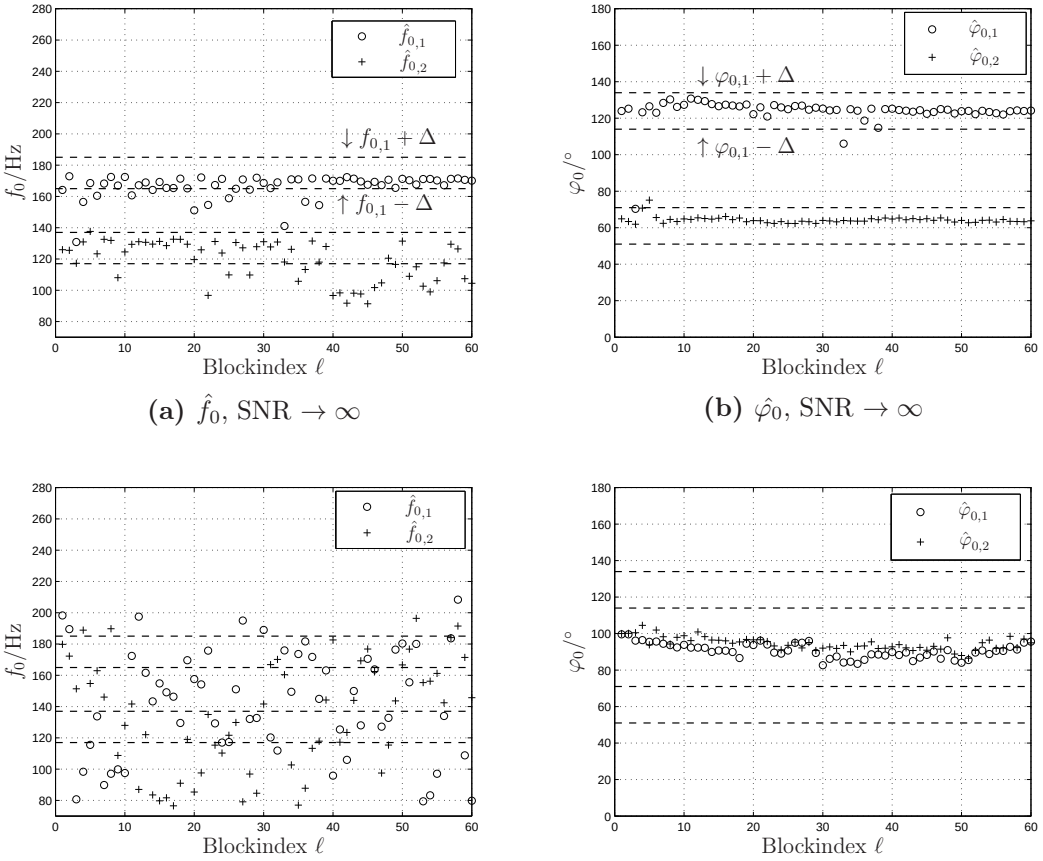
$$Acc_{\varphi_0} = \frac{1}{P}\sum_{\ell=1}^{P}\delta(\hat{\varphi}_0, \varphi_0) \tag{5.7}$$

$$Acc = \frac{1}{P}\sum_{\ell=1}^{P}\delta(\hat{f}_0, f_0) \wedge \delta(\hat{\varphi}_0, \varphi_0) \tag{5.8}$$

$$\delta(a,b) = \begin{cases} 1 & ,\text{wenn} |a-b| \leq \Delta \\ 0 & ,\text{sonst} \end{cases} \tag{5.9}$$

Das *Acc*-Maß beschreibt die relative Trefferrate des Algorithmus über alle berücksichtigten Zeitblöcke P. Wobei die Quelle nur dann als richtig erkannt gilt, wenn die Schätzung innerhalb festgelegter Toleranzgrenzen Δ liegt. Ansonsten wird die Schätzung als falsch verworfen und die Trefferrate sinkt. Es wird zwischen der kombinerten Trefferrate *Acc* und den eindimensionalen Varianten Acc_{f_0}, Acc_{φ_0} unterschieden. Bei der kombinierten Version zählt die Schätzung nur für den Fall als korrekt, wenn die Einfallsrichtung und Grundfrequenz gleichzeitig richtig erkannt werden und ist damit prinzipbedingt geringer als die eindimensionalen Varianten. Als Toleranzgrenzen für die Richtungsschätzung wurden $\pm 10°$ und für die Grundfrequenzschätzung ± 10 Hz gewählt. Jede Simulation wurde fünf mal wiederholt und gemittelt, um den Einfluss der zufälligen Partikelvorhersage auf das Ergebnis zu verringern.

Abbildung 5.4 zeigt die Quellenschätzung aus denen sich die *Acc*-Trefferrate ermittelt, aufgeteilt in die Grundfrequenz- und Richtungsschätzung. Die Toleranzgrenzen sind durch gestrichelte Linien dargestellt. Die Abbildungen zeigen die Schätzungen für zwei simultane Sprecher in zwei unterschiedlich starken Störgeräuschsituationen. Es ist zu erkennen, dass sich die Richtungsschätzung mit zunehmenden Störgeräusch zur Ebenenmitte verschiebt. Die Grundfrequenzschätzung hingegen wird mit zunehmenden Störgeräusch undifferenzierter.



(a) $\hat{f}_0$, SNR $\to \infty$ **(b)** $\hat{\varphi}_0$, SNR $\to \infty$

(c) $\hat{f}_0$, SNR $= -10\,\mathrm{dB}$ **(d)** $\hat{\varphi}_0$, SNR $= -10\,\mathrm{dB}$

Abbildung 5.4: Zeitverlauf der Grundfrequenz- (links) und Richtungsschätzung (rechts) für zwei verschiedene Störgeräuschsituationen bei zwei simultanen Sprechern mit $f_0 = 200\,\mathrm{Hz}, 160\,\mathrm{Hz}$ aus $\varphi_0 = 64°, 121°$. Die Linien deuten den Toleranzbereich für eine korrekte Schätzung in der *Acc*-Bewertung an, $\Delta f_0 = \pm 10\,\mathrm{Hz}$ und $\Delta\varphi_0 = \pm 10°$. Die Blocklänge betrug 85 ms bei $f_s = 48\,\mathrm{kHz}$.

5.3 Partikel-Filter Evaluation

Der Partikel-Filter Algorithmus bietet eine Reihe von Parametern insbesondere bei der physikalisch motivierten Vorhersage der neuen Partikelzustände, aber auch bei der Anzahl der verwendeten Partikel pro Signalquelle (siehe Abschnitt 4.1). Um eine möglichst gute Wahl der Freiheitsgrade zu erhalten, wurden ausgehend von den in [WLW03, HR10] vorgeschlagenen Einstellungen verschiedene Bedingungen untersucht. Die in [WLW03, HR10] beschriebenen Einstellungen lauteten:

$$L = 100, \tag{5.10}$$

$$v = 1\,\mathrm{m/s}, \tag{5.11}$$

$$\beta_{x,y} = 10\,\mathrm{s}^{-1}. \tag{5.12}$$

Die Grundfrequenz f_0 wurde in der Literatur nicht näher berücksichtigt. Die Geschwindigkeit mit der sich die Grundfrequenz ändern kann, wurde daher Anfangs selbst mit

$$\beta_{f_0} = 50\,\mathrm{Hz/s} \tag{5.13}$$

gewählt. Bezogen auf die Blockverarbeitung mit einer Zeitspanne von z.B. 85 ms ($f_s = 48\,\mathrm{kHz}$, $N = 4096$ Samples) ergibt sich daraus eine maximale Grundfrequenzänderung pro Blockverarbeitung von 4,25 Hz. Die Partikel-Evaluation diente vorrangig der Wertebestimmung für die zulässige Grundfrequenzänderung.

Ausgehend von diesen Parametern wurde für ein Testszenario von zwei simultanen Sprechern in moderaten Störgeräusch (SNR = 10 dB) und einer simulierten Nachhallzeit von 100 ms jeweils ein Parameter variiert, während alle restlichen Parameter konstant gehalten wurden. Die sich ergebenen Trefferraten sind in Tabelle 5.2 aufgeführt. Aus der Berechnung hervorgegangene Parameter sind fett gedruckt.

Tabelle 5.2: *Acc*-Trefferraten bei Variation der Parameter des Partikel-Filter unter Verwendung der KLDS-Kernfunktion, $\mathrm{T}_3\{\cdot\}$ Phasentransformation, MC-CC-Kombination und cepstraler Gewichtung sowie Filterbankvorverarbeitung $F = 64$. Die Trefferraten der zwei zu detektierenden Signalquellen sind gemittelt für die Untersuchungen aufgetragen. Moderate akustische Vorraussetzungen von SNR = 10 dB und $\tau_{60} = 100$ ms (simuliert). Line-Array mit 6 Mikrofonen. Kombinierte Grundfrequenz und Richtungsschätzung im Spektralbereich.

	Acc in %				*Acc* in %		
Partikel	φ_0	f_0	$f_0 \wedge \varphi_0$	$\beta_{x,y}\,[\mathrm{s}^{-1}]$	φ_0	f_0	$f_0 \wedge \varphi_0$
10	64,95	52,77	38,31	1	95,24	95,71	91,90
50	98,10	96,67	95,24	5	90,95	95,71	89,52
100	91,90	97,62	90,95	**10**	94,86	96,29	92,29
200	90,48	97,14	89,05	15	93,81	96,67	92,38
300	90,48	97,14	89,05	20	93,81	94,76	90,00
$\beta_{f_0}\,[\mathrm{s}^{-1}]$	φ_0	f_0	$f_0 \wedge \varphi_0$	$v_{x,y}\,[\mathrm{m/s}]$	φ_0	f_0	$f_0 \wedge \varphi_0$
5	87,56	92,88	87,56	0,2	91,43	95,24	89,05
50	93,81	95,24	91,43	0,5	90,95	88,57	85,24
100	79,52	94,29	79,05	**1**	93,81	95,24	91,90
200	91,90	95,24	87,62	1,5	89,05	93,33	86,19
360	91,43	88,10	81,90	2	86,19	95,71	84,29
500	75,71	80,95	58,10				

Die Variation der, die Partikelbewegung beeinflussenden, Parameter $\beta_{x,y}$ und $v_{x,y}$ ergaben in den Untersuchungen (Tabelle 5.2) keine sonderliche Abhängigkeit von dessen Parametergröße für statische Quellen. Daher wurden sie bis auf weiteres auf die in der Literatur [WLW03] beschriebenen Größenwerte $\beta_{x,y} = 10\,\mathrm{s}^{-1}$ und

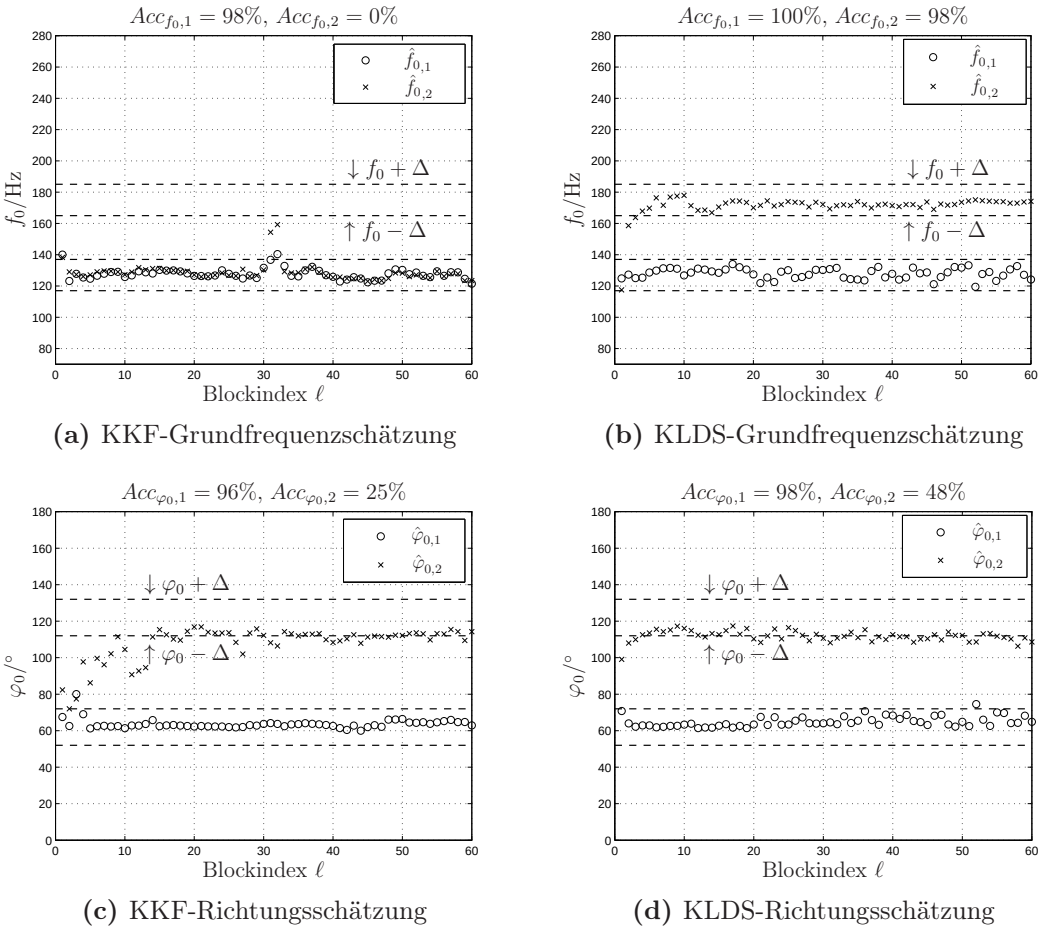
$v_{x,y} = 1\,\mathrm{m/s}$ gesetzt. Die Veränderung der Partikelanzahl ließ bei der beschriebenen Testkonstellation einen Wert von 50 Partikeln pro Quelle favorisieren. Für den Parameter der Grundfrequenzschwankung hat sich ebenso eine Tendenz für eine kleinere Parametergröße gezeigt und wurde somit auf $\beta_{f_0} = 50\,\mathrm{s}^{-1}$ favorisiert.

Alle nachfolgenden Berechnungen wurden, soweit nicht anders angegeben, mit den in Tabelle 5.2 fett gedruckten Werten vorgenommen. Die Breite der Sperr-Filter (siehe Abschnitt 4.2) wurde für die KKF-Variante auf $\beta_{br} = 10$ und für die KLDS-Variante auf $\beta_{br} = 1$ gesetzt. Die Quellenpositionen blieben bei den Berechnungen für die Partikelparameter konstant.

5.4 Vergleich der Kernalgorithmen

In diesem Abschnitt werden die beiden Kernfunktionen der kombinierten Grundfrequenz- und Richtungsschätzung miteinander verglichen. Das sind zum einen das Verfahren im Zeitbereich über die KKF und zum anderen das Verfahren im Frequenzbereich über das KLDS. Dabei wurde besonders darauf Wert gelegt, ob mit dem jeweiligen Algorithmus eine Schätzung mindestens zweier Sprecher möglich ist. Im Zuge dessen wurden beide Verfahren in einer Situation mit zwei simultanen Sprechern (männlich und weiblich), einer simulierten Nachhallzeit von 250 ms und einem SNR von 20 dB analysiert. Es kamen jeweils die Erweiterungen der MCCC-Kombination, cepstralen Gewichtung sowie Filterbankvorverarbeitung $F = 64$ zum Einsatz. Die KLDS-Kernfunktion wurde mit der $\mathrm{T}_3\{\cdot\}$ Phasentransformation verwendet. Es zeigte sich, dass die KKF-Version eine wesentlich schlechtere Erkennerleistung, verglichen mit der KLDS-Version, aufweist. Zur Unterstreichung dieser Aussage sind in Abbildung 5.5 die Schätzungen beider Verfahren, über die Zeit getrennt für die Grundfrequenzen und Einfallsrichtungen, aufgetragen. Über den Abbildungen sind die jeweilig erzielten Erkennerraten notiert. Beim Vergleich der Abbildungen ist auszumachen, dass die KKF-Variante vorallem in der Grundfrequenzschätzung schlechtere Ergebnisse erziehlte. In diesem Fall wurde nur eine Grundfrequenz erkannt und fälschlicherweise für beide Quellenschätzungen als Grundfrequenz ausgegeben. Aber auch bei der Schätzung der Einfallsrichtungen weißt die KKF-Version eine geringere Erkennerrate auf.

Als Ursache für das schlechtere Ergebnis der KKF-Version konnte das Verfahren der Sperr-Filterung ausgemacht werden. Im Fall der Berechnung im Zeitbereich (KKF) muss für eine korrekte Sperr-Filterung der Einfallswinkel und die Grundfrequenz richtig geschätzt werden. Wohingegen für die Sperr-Filterung im Frequenzbereich alleinig die Grundfrequenzschätzungen für die Filterbildung ausschlaggebend ist. Die Resultate der Sperr-Filter sind in Abbildung 5.6 beispielhaft dargestellt. Auf der linken Seite ist die KKF-Sperr-Filterung abgebildet. Hierbei zeigt sich, wenn mit einer falschen Grundfrequenzschätzung jedoch korrekter DoA-Schätzung eine Sperr-Filterung durchgeführt wird, so wird die korrekte Grundfrequenz ebenfalls stark unterdrückt. Dieser Sachverhalt deutet auf eine hohe Bedeutung des ersten Maxima in der Kreuzkorrelation hin.



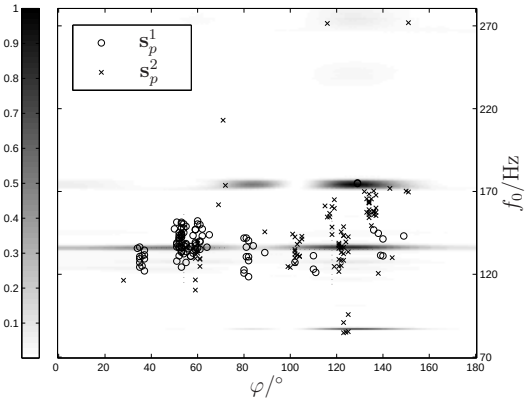
(a) KKF-Grundfrequenzschätzung **(b)** KLDS-Grundfrequenzschätzung

(c) KKF-Richtungsschätzung **(d)** KLDS-Richtungsschätzung

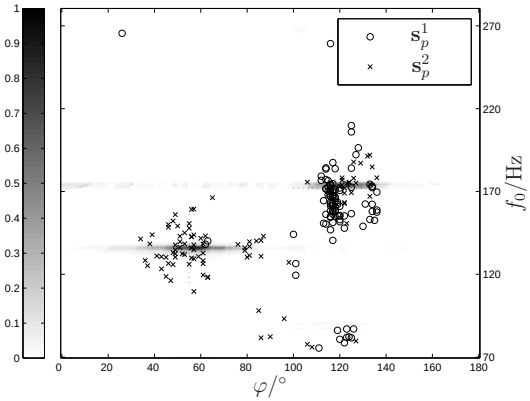
Abbildung 5.5: Grundfrequenz- und Richtungsschätzungen dargestellt über aufeinanderfolgend, verarbeitete Zeitblöcke ℓ. Auf der linken Seit sind die Resultate der KKF-Version und auf der rechten Seite die der KLDS-Version aufgetragen. Die obere Zeile illustriert die Grundfrequenzschätzung. Die untere Zeile stellt die Richtungsschätzung dar.

Bei der Sperr-Filterung im Frequenzbereich (Betragsunterdrückung) wird mit korrekter Grundfrequenzschätzung und geeignet gewählter Phasentransformation zwangsläufig die korrekte Einfallsrichtung unterdrückt, da für die Frequenzdarstellung pro Frequenzbin nur eine Phaseninformation zur Verfügung steht. Es können somit bei der KLDS-Berechnung nur dann zwei unterschiedliche simultane Quellen ausgemacht werden, wenn diese auch unterschiedliche Grundfrequenzen besitzen. Eine erfolgreiche Sperr-Filterung ist in Abbildung 5.6 dargestellt.

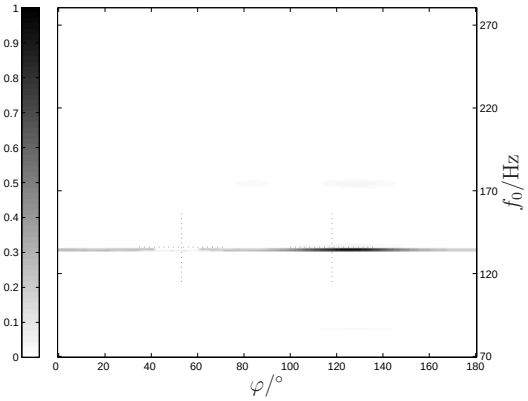
Als Konsequenz aus diesen Erkenntnissen wurde für die nachfolgenden Betrachtungen nur noch die KLDS-Berechnungsmethode angewandt.



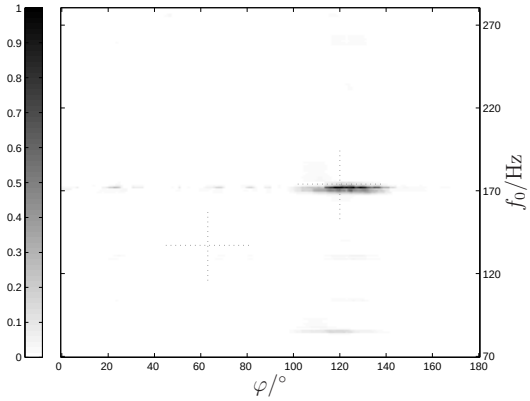
(a) Ergebnisebene mit Partikelzuständen nach 12 Iterationen.



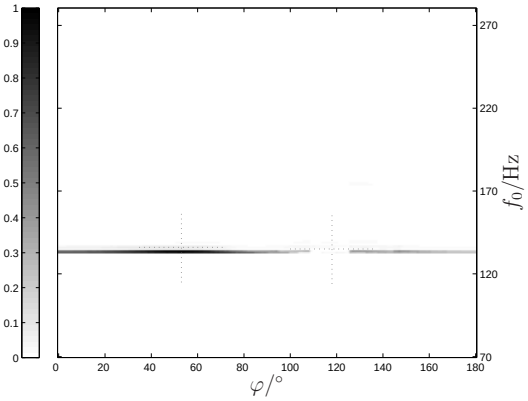
(b) Ergebnisebene mit Partikelzuständen nach 12 Iterationen.



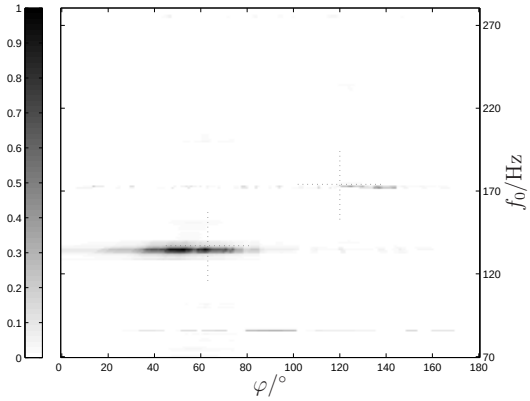
(c) Sperr-Filterung für Schätzung 1.



(d) Sperr-Filterung für Schätzung 1.



(e) Sperr-Filterung für Schätzung 2.



(f) Sperr-Filterung für Schätzung 2.

Abbildung 5.6: Ergebnisebene der kombinierten Schätzung. Linke Seite: KKF-Berechnung. Rechte Seite: KLDS-Berechnung. Erste Zeile: Partikelzustände zur Schätzung. Zweite und dritte Zeile: Resultate nach Sperr-Filterung für jeweils eine Quellenschätzung.

5.5 Vergleich der Phasentransformationen

Wie in Abschnitt 3.3 beschrieben, wurden für die kombinierte Schätzung über das KLDS drei unterschiedliche Phasentransformationen implementiert, die es an dieser Stelle zu evaluieren galt. Im Rahmen dessen wurden mit allen drei Varianten abermals eine und dieselbe akustische Szene analysiert. In diesem Fall wurden zwei simultane Sprecher gefaltet mit den realen Impulsantworten und 10 dB SNR als Vorlage ausgewählt. Die mit allen drei Phasentransformationen erzielten Erkennerraten sind in Abbildung 5.7 dargestellt. Für jede Quelle wird die Erkennerrate nach Abschnitt 5.2 unterschieden. Es lässt sich feststellen, dass sich eine breitere Vorzugsrichtung positiv auf die mit dem System erzielte Erkennerrate auswirkt. Die $\mathrm{T}_1\{\cdot\}$ zeigte auch ohne justierbare Vorzugsrichtung bereits eine unerwartet hohe Erkennerleistung. Die beiden weiteren Phasentransformationen $\mathrm{T}_2\{\cdot\}$ und $\mathrm{T}_3\{\cdot\}$ wurden für die in Abbildung 5.7 erzielten Trefferraten nur mit relativ großen Koeffizienten von $\beta = 0{,}7$ erzielt.

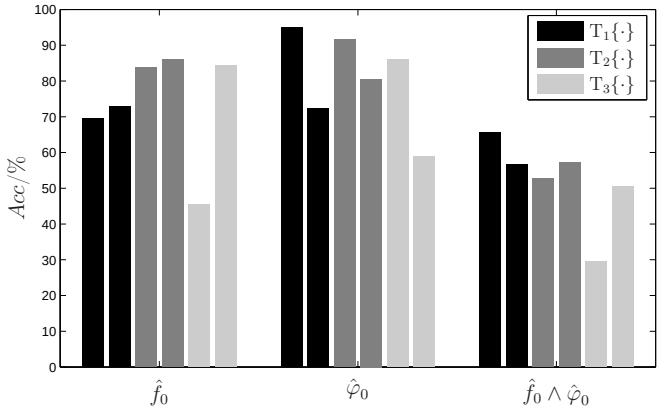
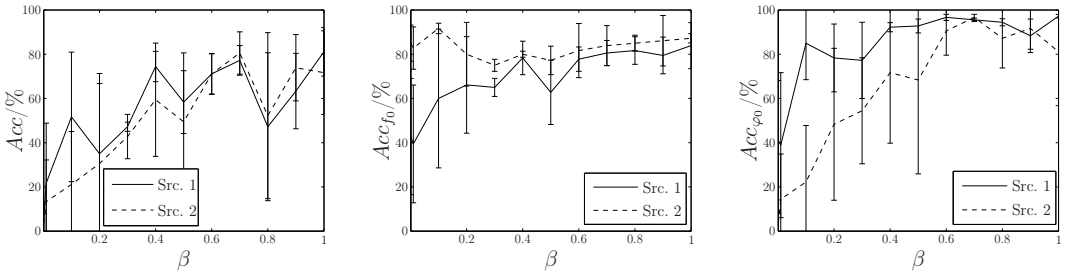


Abbildung 5.7: Trefferraten der Phasentransformationen unterteilt in die Trefferraten für Grundfrequenz, Einfallsrichtung und kombinerte Trefferrate, getrennt für jede einzelne Quelle. Als Vorlage dienten zwei simultane Sprecher (männlich und weiblich) gefaltet mit realen Impulsantworten (550 ms) bei 10 dB SNR.

Als Favorit der Phasentransformation wurde die Variante $\mathrm{T}_2\{\cdot\}$ ausgewählt. Im Gegensatz zur einfachen $\mathrm{T}_1\{\cdot\}$ Transformation bietet sie mit dem Parameter β noch eine Anpassmöglichkeit und wurde daher bevorzugt. Für die favorisierte Phasentransformation $\mathrm{T}_2\{\cdot\}$ sind in Abbildung 5.8 die *Acc*-Trefferraten für unterschiedliche β abgebildet. Anhand dieser Grafik wurde der Parameter β auf einen Wert von 0,7 eingestellt.



(a) Komb. Trefferrate *Acc*. **(b)** Grundfrequenz-Trefferrate Acc_{f_0}. **(c)** DoA-Trefferrate Acc_{φ_0}.

Abbildung 5.8: *Acc*-Trefferraten für unterschiedliche β bei Phasentransformation $\mathrm{T}_2\{\cdot\}$.

Die selbst entworfene Phasentransformation $\mathrm{T}_3\{\cdot\}$ konnte nicht ganz an die Erkennerraten der übrigen Transformationen anknüpfen. Als Ursache für dieses Resultat ist die Winkelauflösung der Phasentransformation zu nennen. Im Rahmen dieser Arbeit wurde eine Winkelauflösung von einem Grad gewählt, jedoch zeigte besonders die selbst konzipierte Phasentransformation $\mathrm{T}_3\{\cdot\}$ eine starke Abhängigkeit von der Übereinstimmung der berechneten Phase mit der gemessenen Phase. Diesen Nachteil können die Funktionen $\mathrm{T}_1\{\cdot\}$ und $\mathrm{T}_2\{\cdot\}$ über ihre kontinuierlichen Gewichtungspattern einhergend mit der bei $\mathrm{T}_2\{\cdot\}$ groß gewählten Vorzugsrichtung kompensieren (siehe auch Abschnitt 3.3).

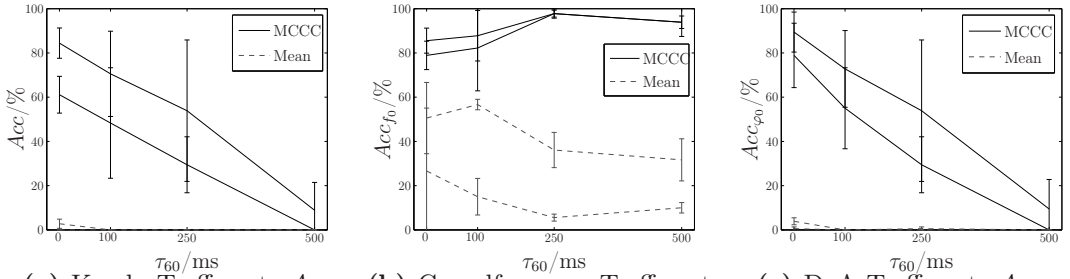
5.6 Evaluation der Erweiterungen

Neben den bis hierher ausgewerteten Verfahren und Parametern galt es auch die auf den Kernalgorithmen aufbauenden Erweiterungen zu evaluieren. Somit werden in diesem Abschnitt die eigenen Erweiterungen der MCCC-Kombination von mehreren Mikrofonpaaren (vgl. Abschnitt 3.4) und die Phat-Gewichtung (vgl. Abschnitt 3.7) aber auch die aus der Literatur entnommenen Verfahren der Filterbankvorverarbeitung (vgl. Abschnitt 3.6) und Cepstrum-Gewichtung (vgl. Abschnitt 3.5) behandelt. In diesem Abschnitt beruhen die Beurteilungen auf der Variation der Nachhallzeiten (simulierte RIR) bei einem SNR von 10 dB für abermals zwei simultane Sprecher (männlich und weiblich). In den Abbildungen werden immer jeweils zwei Verfahren gegenübergestellt, die sich jeweils nur um die untersuchte Erweiterung unterscheiden.

5.6.1 MCCC-Kombination

Beginnend mit Abbildung 5.9 wurden die Auswirkungen der MCCC-Kombination gegenüber einer einfachen Mittelung der Mikrofonpaare beurteilt. Wie man aus den Graphen entnehmen kann, ist die MCCC-Kombination in allen Belangen

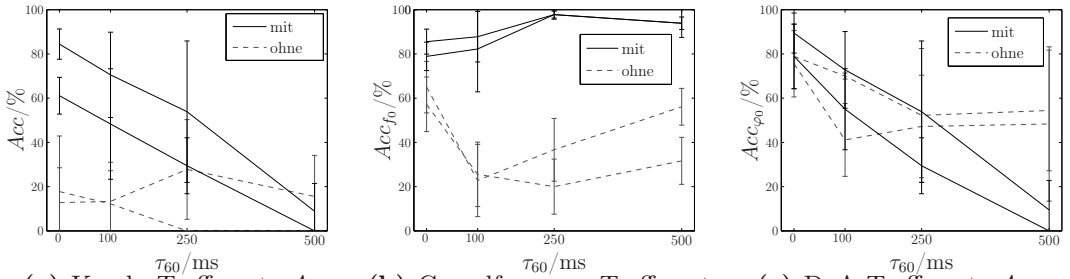
der einfachen Mittelung überlegen. *Man kann sogar sagen, dass bei der gewählten Line-Array-Geometrie und Störgeräuschsituation ohne die MCCC-Kombination keine kombinierte Schätzung mit dem untersuchten Verfahren möglich ist.* Die aus der einfachen Mittelung resultierenden, über die Grundfrequenz- und DoA-Schätzung aufgespannten, Ergebnisebenen weisen einen deutlich stärkeren, rauschhaften Charakter auf als die über die MCCC-Kombination generierten Ergebnisebenen. Als Ursache dafür, lässt sich die Begründung aus Abschnitt 3.4 heranziehen. So dass bei der MCCC-Berechnung nur dann eine hohe Schätzung erzielt wird, wenn alle Mikrofonpaare ein hohes Resultat implizieren. Wohingegen sich bei der einfachen Mittelung Fehler auch gegenseitig ausgleichen können.



(a) Komb. Trefferrate *Acc*. **(b)** Grundfrequenz-Trefferrate Acc_{f_0}. **(c)** DoA-Trefferrate Acc_{φ_0}.

Abbildung 5.9: Vergleich der MCCC-Kombination mit einfacher Mittelung bei unterschiedlichen Nachhallzeiten.

5.6.2 Filterbankvorverarbeitung



(a) Komb. Trefferrate *Acc*. **(b)** Grundfrequenz-Trefferrate Acc_{f_0}. **(c)** DoA-Trefferrate Acc_{φ_0}.

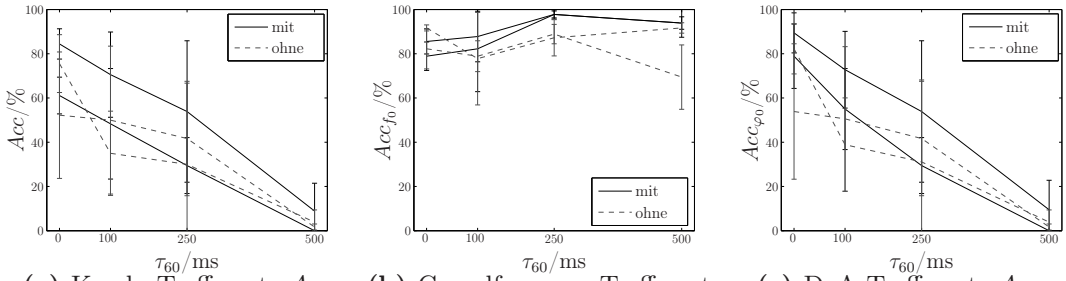
Abbildung 5.10: Vergleich der Trefferraten mit sowie ohne Filterbankvorverarbeitung bei unterschiedlichen Nachhallzeiten.

In Abbildung 5.10 sind die Auswirkungen der Filterbankvorverarbeitung dokumentiert. Wie bereits in Abschnitt 3.6 erwartet, führt die Unterteilung des Spektrums in mehrere bandpassgewichtete Kanäle zu einer Verbesserung der Grundfrequenzschätzung bei simultanen Sprechern. Begründet wurde dies durch die ge-

trennte Verteilung der Nutzsignalenerige in unterschiedliche Spektren. Die Schätzung der Einfallsrichtung scheint bei größeren Nachhallzeiten ($\geq 250\,\text{ms}$) jedoch unter der Filterbankvorverarbeitung an Genauigkeit zu verlieren. Dieses Phänomen berücksichtigend, befasst sich der letzte Vergleich in diesem Abschnitt nochmals mit der Filterbankvorverarbeitung unter optimaler Kombination aller übrigen Erweiterungen.

5.6.3 Cepstrum Gewichtung

Aus der Abbildung 5.11 ist der Einfluss der Cepstrum-Gewichtung zu entnehmen. Es zeigt sich eine leichte Verbesserung sowohl in der Grundfrequenz- als auch Richtungsschätzung. Diese ist zwar nicht so deutlich ausgeprägt wie bei den zurvor gezeigten Vergleichen, jedoch führt sie noch immer zu einem Anstieg der *Acc*-Rate bei einem überwiegenden Teil der Simulationen. Die erste Quelle liegt mit ihren Ergebnissen immer überhalb der Simulation ohne Cepstrum-Gewichtung. Bei der zweite Quelle liegt die Ergebnisse teilweise mit auch unterhalb der Simulationen ohne Cepstrum-Gewichung. Die Cepstrum-Gewichtung soll theoretisch vornehmlich die Grundfrequenzschätzung verbessern, hat aber auch Einfluss auf die DoA-Schätzung. Das Prinzip der einheitlichen Grundfrequenzgewichtung für alle Einfallsrichtungen beeinflusst die Gesamtschätzung positiv.



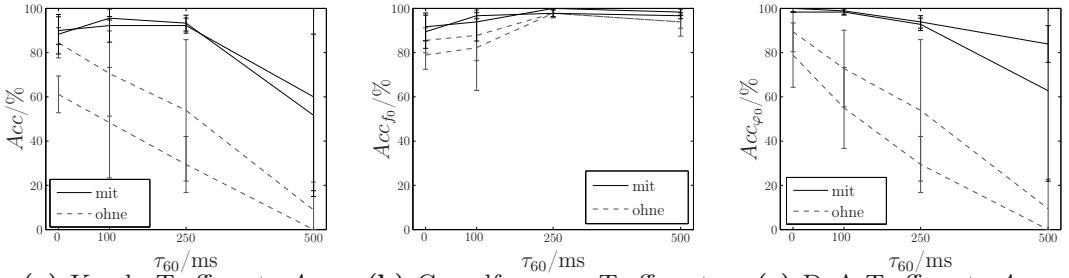
(a) Komb. Trefferrate *Acc*. **(b)** Grundfrequenz-Trefferrate Acc_{f_0}. **(c)** DoA-Trefferrate Acc_{φ_0}.

Abbildung 5.11: *Acc*-Trefferraten mit sowie ohne Cepstrum-Gewichtung bei unterschiedlichen Nachhallzeiten.

5.6.4 GCC-Phat Gewichtung

Dieser Vergleich befasst sich mit der selbst entworfenen GCC-Phat-Gewichtung. Im Gegensatz zur Cepstrum-Gewichtung ist diese für alle Grundfrequenzen der Ergebnisebene identisch, führt aber eine Gewichtung der Einfallsrichtungen aus. Wie aus Abbildung 5.12 zu entnehmen ist, resultiert daher insbesondere in der Richtungschätzung eine deutliche Steigerung der *Acc*-Trefferrate. Ein Vergleich der an dieser Stelle erzeugten Ergebnisebenen ermöglicht eine Begründung der Verbesserung. Ohne GCC-Phat-Gewichtung erscheinen mögliche Quellenschätzungen, bezogen auf die Einfallsrichtung, durch recht breite Schätzungen. Eben

dieses Manko kann durch die GCC-Phat-Gewichtung reduziert werden. So dass an dieser Stelle auch erstmals eine hohe (für $\tau_{60} \leq 250\,\text{ms}$) kombinierte Trefferate zu verzeichnen ist.

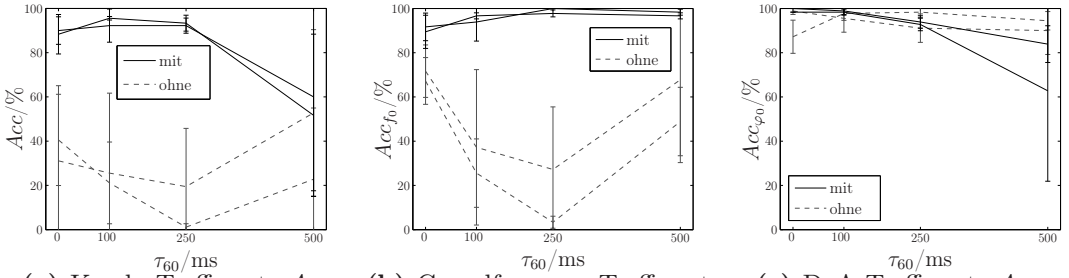


(a) Komb. Trefferrate *Acc*. **(b)** Grundfrequenz-Trefferrate Acc_{f_0}. **(c)** DoA-Trefferrate Acc_{φ_0}.

Abbildung 5.12: *Acc*-Trefferraten mit sowie ohne GCC-Phat-Gewichtung bei unterschiedlichen Nachhallzeiten.

5.6.5 Optimale Kombination

Bezug nehmend auf die Evaluationsergebnisse mit bzw. ohne Filterbankvorverarbeitung wird nun zu guter Letzt nocheinmal mit zusätzlicher Cepstrum- und GCC-Phat-Gewichtung der Einfluss der Filterbankvorverarbeitung untersucht. Anhand der in Abbildung 5.13 gezeigten Graphen lässt sich an dieser Stelle feststellen, dass die Filterbankvorverarbeitung einen sinnvollen Nutzen für die kombinierte Schätzung erbringt. Durch sie ist eine taugliche Grundfrequenzschätzung gewährleistet, wohingegen die Richtungschätzung nur unwesentlich verschlechtert wird.



(a) Komb. Trefferrate *Acc*. **(b)** Grundfrequenz-Trefferrate Acc_{f_0}. **(c)** DoA-Trefferrate Acc_{φ_0}.

Abbildung 5.13: Nochmalige Kontrolle der Filterbankvorverarbeitung. Diesmal jedoch mit GCC-Phat-Gewichtung bei unterschiedlichen Nachhallzeiten.

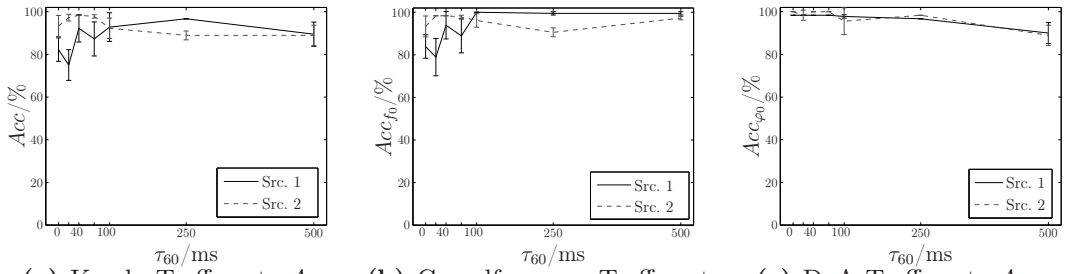
Zusammenfassend besteht die kombinierte Schätzung in den folgenden Untersuchungen jeweils aus der KLDS-Kernfunktion mit MCCC-Kombination, Cepstrum-Gewichtung, Filterbankvorverarbeitung und GCC-Phat-Gewichtung.

5.7 Nachhall- und Störgeräuscheinfluss

Nachdem nun die einzelnen Parameter und Erweiterungen in ihrem Einfluss untersucht wurden sind, soll in diesem Kapitel geklärt werden, welche Erkennerleistung durch die favorisierte Kombination erreicht werden kann. Dazu wurde die bereits des öfteren beschriebene Situation zweier simultaner Sprecher mit $f_{0,1} = 175\,\mathrm{Hz}$ und $f_{0,2} = 127\,\mathrm{Hz}$ aus $\varphi_{0,1} = 64°$ und $\varphi_{0,2} = 127°$ abermals mit dem nun optimierten Verfahren untersucht. Die Auswertung erfolgte einerseits für unterschiedliche Nachhallzeiten ohne zusätzliches diffuses Rauschen. Andererseits wurden verschiedene Störgeräuschpegel simuliert, diese jedoch wiederum ohne Nachhall präsentiert. Zuletzt erfolgte noch eine Auswertung mit realen RIR bei verschiedenen Störgeräuschpegeln.

An dieser Stelle seien nocheinmal alle für diesen Abschnitt relevanten Parameter und Algorithmen zur Übersicht aufgelistet:

- KLDS-Kernfunktion mit $\mathrm{T}_2\{\cdot\}$ Phasentransformation
- MCCC-Kombination,
- GCC-Phat-Gewichtung,
- Cepstrum-Gewichtung,
- Filterbankvorverarbeitung $F = 64$,
- Partikelfilter mit 50 Elementen,
- Langevin-Modell mit $\beta_{x,y} = 10\,\mathrm{s}^{-1}$, $v_{x,y} = 1\,\mathrm{m/s}$, $\beta_{f_0} = 50\,\mathrm{s}^{-1}$



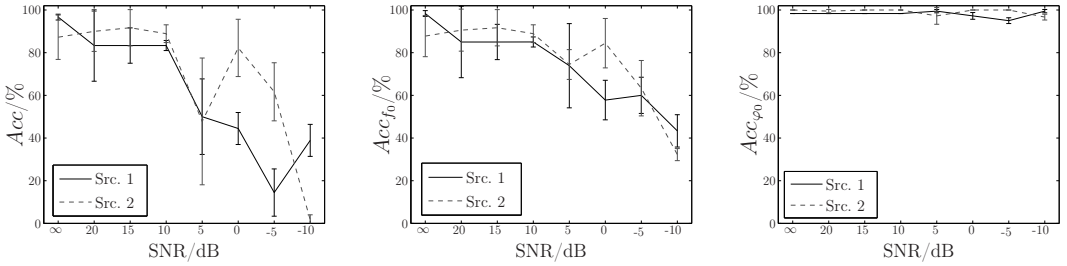
(a) Komb. Trefferrate *Acc*. **(b)** Grundfrequenz-Trefferrate Acc_{f_0}. **(c)** DoA-Trefferrate Acc_{φ_0}.

Abbildung 5.14: *Acc*-Trefferraten für unterschiedliche Nachhallzeiten, ohne Störgeräusch bei zwei simultanen Sprechern.

Beginnend mit Abbildung 5.14 sind die Ergebnisse für unterschiedliche Nachhallzeiten (simuliert) ohne Störgeräuschbeeinflussung aufgetragen. Bei der Richtungsschätzung lässt sich eine durchweg hohe Erkennerleistung feststellen, die erwartungsgemäß zu längeren Nachhallzeiten leicht abnimmt. Auch die Grundfrequenzschätzung zeichnet sich durch eine hohe Trefferrate für alle Nachhallzeiten aus. Die bei kurzen Nachhallzeiten ersichtlichen geringen Einbrüche wurden nicht

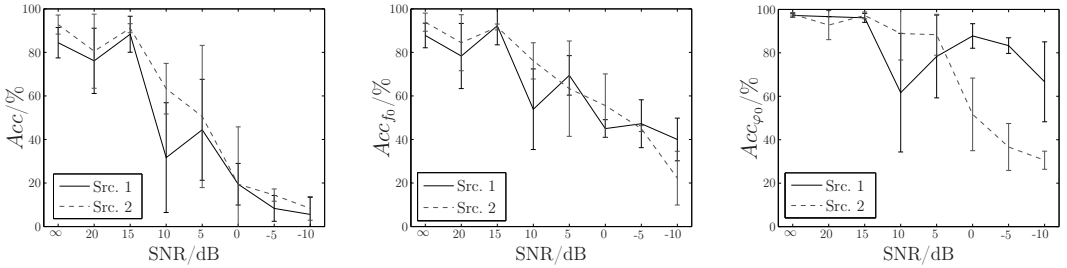
weiter untersucht. Alles zusammen ergibt dies eine durchschnittliche kombinierte Trefferrate von ca. 90 %.

Bei der Betrachtung der Ergebnisse für verschiedene Störgeräuschpegel zeigt sich eine differenzierte Sensitivität des Algorithmus. Aus Abbildung 5.15 lässt sich entnehmen, dass die Richtungsschätzung wesentlich unsensibler auf kleinere SNR-Werte reagiert als die Grundfrequenzschätzung. Die Grundfrequenz-Trefferrate nimmt mit zunehmendem Störgeräuschpegel deutlich ab, wohingegen die Richtungschätzung keine merklichen Einbußen im betrachteten SNR-Bereich zu verzeichnen hat. Beeinflusst von der Grundfrequenzschätzung zeigt auch die kombinierte Trefferrate zu höheren Rauschpegeln deutliche Einbußen. In der Summe kann aber festgehalten werden, dass bis zu einem SNR von 15 dB noch sehr wohl gute kombinierte Trefferraten zu erzielen sind.



(a) Komb. Trefferrate *Acc*. **(b)** Grundfrequenz-Trefferrate Acc_{f_0}. **(c)** DoA-Trefferrate Acc_{φ_0}.

Abbildung 5.15: *Acc*-Trefferraten für unterschiedliche SNR, ohne Nachhallzeit bei zwei simultanen Sprechern.



(a) Komb. Trefferrate *Acc*. **(b)** Grundfrequenz-Trefferrate Acc_{f_0}. **(c)** DoA-Trefferrate Acc_{φ_0}.

Abbildung 5.16: *Acc*-Trefferraten für unterschiedliche SNR, reale Impulsantworten bei zwei simultanen Sprechern.

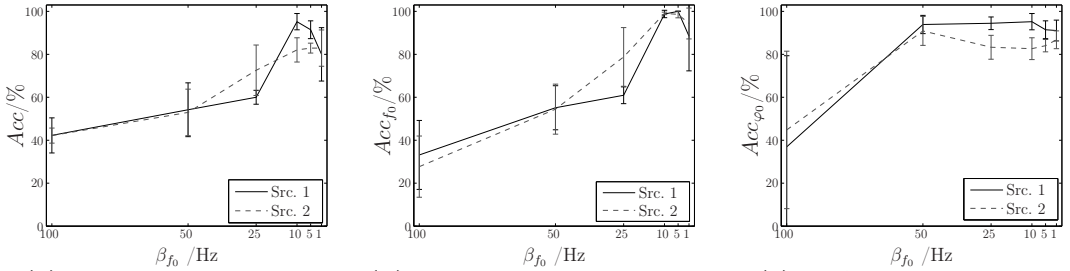
Im Interesse, das untersuchte Verfahren auch unter realen Umweltbedingungen einsetzen zu können und durch die motivierenden Trefferraten bei idealisierten Bedingungen, sind in Abbildung 5.16 Trefferraten bei realen RIR und unterschiedlichen Störgeräuschpegeln dargestellt. Bis zu einem Pegel von 15 dB SNR sind recht gute Trefferraten auszumachen. Bei zunehmenden Störgeräusch sinkt

die Erkennerleistung sowohl bei der Grundfrequenz- als auch Richtungsschätzung. Es lässt sich der Schluss ziehen, dass das Verfahren bei guten SNR-Werten sowie moderaten Nachhallzeiten sehr wohl unter realen Bedingungen einsetzbar ist.

5.8 Kombinierte Schätzung bei sich bewegenden Quellen

Bis zu diesem Abschnitt wurde für die Auswertung immer von statischen Quellen ausgegangen. Allerdings ist es sehr wohl möglich, dass eine Person, während sie spricht, sich auch im Raum bewegt. Genau diesen Sachverhalt versucht die kombinierte Grundfrequenz- und Richtungsschätzung mit Partikel-Filterung zu berücksichtigen. Zum einen wurde eben bei der Auswertung der Quellencharakerisik die Grundfrequenz zur Quellenbestimmung immer mit berücksichtigt. Zum anderen bietet das Partikel-Filter über das enthaltene physikalische Modell die Eigenschaft, die Bewegung einer Quelle nachzuvollziehen. Auf Grund dessen wird in diesem Abschnitt die Fähigkeit des beschriebenen Algorithmus untersucht, sich auf bewegende Quellen einzustellen. Im Zuge dessen wurde das Verfahren zunächst mit einer sich bewegenden Quelle evaluiert. Anschließend wurden zwei sich simultan bewegende Quellen verwendet. Als schwierigstes Bedingung wurden zum Abschluss zwei sich räumlich kreuzende simultane Quellen verwendet und überprüft, in wie weit der Algortithmus dies korrekt geschätzt hat. Es wurde untersucht, bis zu welchem Störgeräuscheinfluss ein veritables Tracking möglich ist. Um eine Quellenbewegung zu simulieren wurden die in Abschnitt 5.1 beschriebenen Sprecherpositionen sukzessive als Signalquellen verwendet.

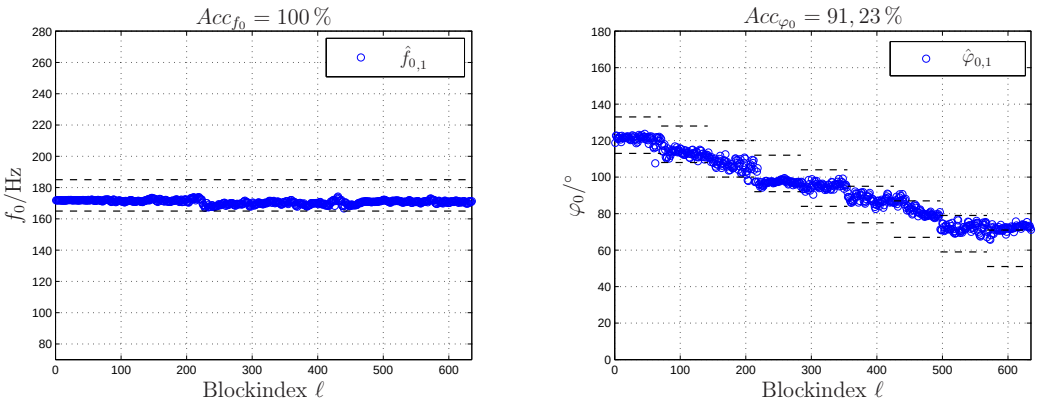
Es stellte sich herraus, dass bei sich bewegenden Quellen für den Parameter β_{f_0} in der Partikelvorhersage abweichend zu Tabelle 5.2 kleinere Werte zu bevorzugen sind. Folglich wurde an dieser Stelle für $\beta_{f_0} = 5$ gewählt. Eine Auswahl verschiedener Parametergrößen und der daraus resultierenen *Acc*-Trefferraten ist in Abbildung 5.17 dargestellt. Besonders bei dicht aneinander positionierten Sprechern führt ein kleinerer Parameterwert zu besseren Resultaten. Für Abbildung 5.17 bewegten sich die Sprecher ausgehend von den Positionen aus Abschnitt 5.3 aufeinander zu (siehe Abbildung 5.19). Im Falle dicht beinanderliegender Sprecherpositionen führen hohe Werte des Parameters β_{f_0} schnell zu Grundfrequenzverwechselungen zwischen den Quellen.



(a) Komb. Trefferrate *Acc*. **(b)** Grundfrequenz-Trefferrate Acc_{f_0}. **(c)** DoA-Trefferrate Acc_{φ_0}.

Abbildung 5.17: *Acc*-Trefferraten für unterschiedliche β_{f_0}, reale Impulsantworten, 15 dB SNR und zwei simultanen, sich aufeinander zubewegenden Sprechern.

In Abbildung 5.18 sind die Quellenschätzungen für eine einzelne sich bewegende Quelle aufgetragen. Wie aus vorhergehenden Kapiteln zu vermuten war, ermöglicht das untersuchte Verfahren eine sehr robuste Schätzung bei nur einer zu verfolgenden Quelle. Der Störgeräuschpegel in Abbildung 5.18 beträgt SNR = 15 dB. Die Grundfrequenz wird ausnahmslos korrekt geschätzt. Bei der Richtungsschätzung sind nur geringe, vernachlässigbare Abweichungen von der wahren Position festzustellen. Aber auch bei größeren Störgeräuschpegeln ist für diese Konstellation noch eine praktikable Erkennerleistung feststellbar. Für weitere *Acc*-Trefferraten bezogen auf die Richtungsschätzung bei verschiedenen SNR sei auf Abbildung 5.22 verwiesen.

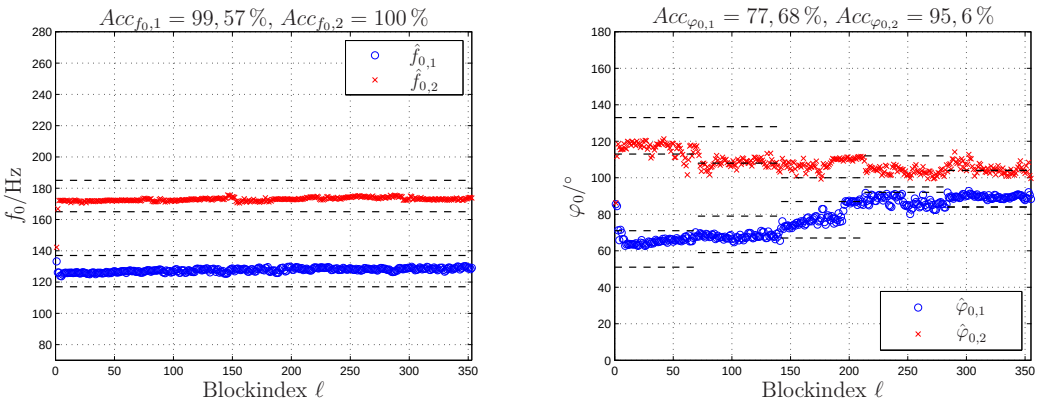


(a) Grundfrequenzschätzung $\hat{f}_0$. **(b)** Einfallsrichtungsschätzung $\hat{\varphi}_0$.

Abbildung 5.18: Kombinierte Quellenschätzung bei einer sich bewegenden Quelle über die Verarbeitungsblöcke ℓ. SNR = 15 dB und reale RIR (τ_{60} = 550 ms).

In der nächsten Stufe wurde versucht, zwei sich simultan aufeinander zubewegende Sprecher zu verfolgen. Abbildung 5.19 zeigt das sich einstellende Resultat bei einem SNR von 15 dB. Auch bei dieser Konstellation wurde die Grundfrequenz zu

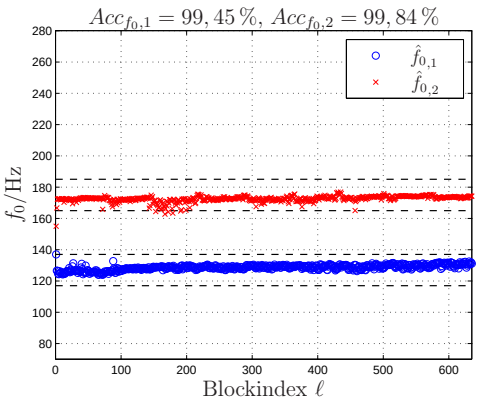
jeder Zeit korrekt geschätzt. Bei der Richtungsschätzung ist der Positionsverlauf des jeweiligen Sprechers eindeutig auszumachen, jedoch sind für einen der beiden Sprecher leichte Abweichungen von der realen Einfallsrichtung zu erkennen. Im Falle das beide Sprecher an der identischen Position (Blockindizees 280-350) stehen, wurden sie trotzdem leicht räumlich getrennt geschätzt.



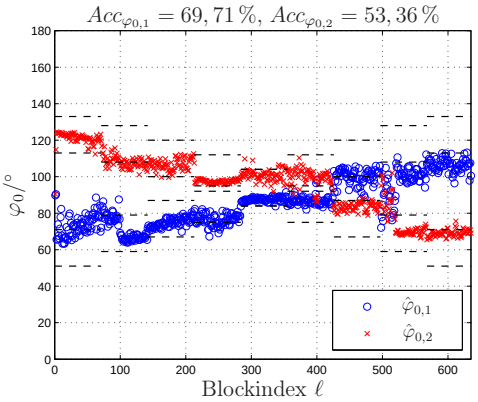
(a) Grundfrequenzschätzung $\hat{f}_0$. **(b)** Einfallsrichtungsschätzung $\hat{\varphi}_0$.

Abbildung 5.19: Kombinierte Quellenschätzung bei zwei sich aufeinander zubewegenden Quellen über die Verarbeitungsblöcke ℓ. SNR = 15 dB und reale RIR ($\tau_{60} = 550\,\text{ms}$).

Im dritten und anspruchvollsten Fall wurden zwei sich kreuzende Quellen simuliert. Die Schwierigkeit bei solch einem Szenario besteht darin, dass die Quellen nach dem Kreuzen weiterhin korrekt zugewiesen werden und in ihrem zeitlichen Verlauf nicht vertauscht werden. Um dies zu gewährleisten, wurde die Grundfrequenzschätzung als zweite Quellencharakteristik zur Differenzierung der Sprecher herangezogen. Abbildung 5.20 illustriert einen korrekten Schätzverlauf für sich kreuzende Quellen bei einem SNR von 15 dB. Die Grundfrequenzen konnten auch hier zu jeder Zeit korrekt geschätzt werden. Der zeitliche Verlauf der Einfallsrichtungen lässt das Kreuzen der Quellen deutlich sichtbar werden. Es ist jedoch eine erneute Verringerung der *Acc*-Trefferraten auszumachen. Dies liegt nicht zu letzt an der schwierigen Schätzung der Einfallsrichtung bei sehr dicht beieinander liegenden Quellen, wie bereits im vorhergehenden Experiment gezeigt wurde. Im Bereich der Verarbeitungsblöcke ca. 360-430 haben sich die Quellen in Wirklichkeit bereits gekreuzt, wurden aber noch immer als unmittelbar nebeneinander liegend geschätzt. Erst bei noch weiter auseinanderliegenden Quellenpositionen wurde das Kreuzen der Quellen erkannt.



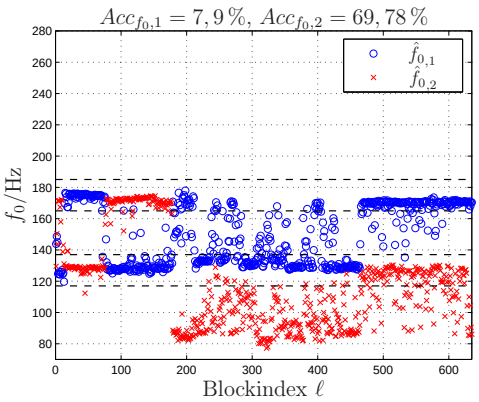
(a) Grundfrequenzschätzung $\hat{f}_0$.



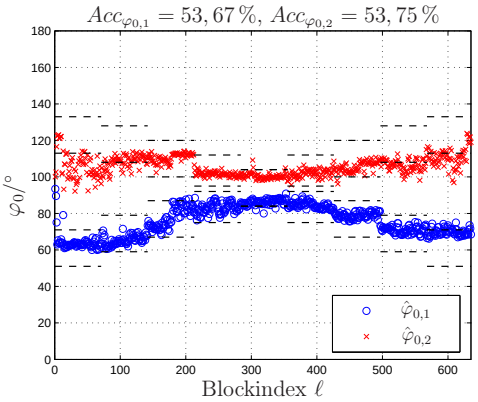
(b) Einfallsrichtungsschätzung $\hat{\varphi}_0$.

Abbildung 5.20: Kombinierte Quellenschätzung bei zwei sich bewegenden und kreuzenden Quellen über die Verarbeitungsblöcke ℓ. SNR = 15 dB und reale RIR ($\tau_{60} = 550$ ms).

Das gleiche Szenario wurde für Abbildung 5.21 nochmals mit einem um 5 dB erhöhten Störgeräuschpegel (SNR = 10 dB) berechnet. In diesem Beispiel konnte das Verfahren den Quellenverlauf nicht mehr korrekt verfolgen. Die Ursache ist dabei klar bei der Grundfrequenzschätzung zu suchen, was sich auch in den *Acc*-Trefferraten äußert. Speziell für die Quelle 2 konnte keine sinnvolle Schätzung der Grundfrequenzen erzielt werden. Dementsprechend resultiert eine fehlerhafte Richtungsschätzung. Zu Beginn genügte eine räumliche Trennung der Quellen, um deren Einfallsrichtung richtig zu ermitteln. Ab Blockindex 350 vertauschte der Algorithmus jedoch den zeitlichen Verlauf der Quellen, da es bei dicht beieinander liegenden Einfallsrichtungen nicht gelang diese, über die Grundfrequenz korrekt zu separieren.



(a) Grundfrequenzschätzung $\hat{f}_0$.

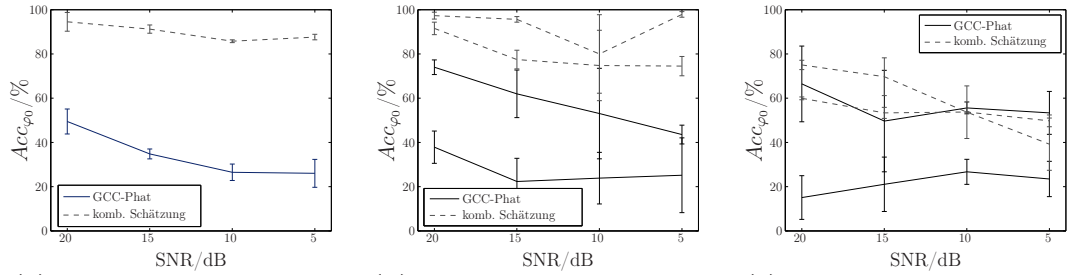


(b) Einfallsrichtungsschätzung $\hat{\varphi}_0$.

Abbildung 5.21: Kombinierte Quellenschätzung bei zwei sich bewegenden und kreuzenden Quellen über die Verarbeitungsblöcke ℓ. SNR = 10 dB und reale RIR ($\tau_{60} = 550$ ms).

5.9 Vergleich mit GCC-Phat-Verfahren

Thema dieses Abschnitts ist der Vergleich der kombinierten Schätzung mit dem wohl bekannten GCC-Phat-Verfahren (Abschnitt 2.3.6), welches in abgewandelter Form bereits als Gewichtungsfunktion (vgl. Abschnitt 3.7) Einzug in das eigene Verfahren gefunden hat. Bei dem Vergleich wurde immer das Partikel-Filter mit anschließender Sperr-Filterung angewendet. Als Pseudo-Likelihood-Funktion $\mathrm{F}(\rho^i, \mathbf{s}_p^i)$ des Partikel-Filters diente jedoch jeweils die kombinierte Schätzung oder das GCC-Phat-Verfahren. Im zweiten Fall wurden das Sperr-Filter und die parametrische Abtastung auf ein einziges notwendiges Maximum reduziert. Denn das GCC-Phat-Verfahren ermöglicht nur eine Richtungsschätzung und beinhaltet keine Auswertung der harmonischen Struktur des Signals über eine Kammabtastung (vgl. Abschnitt 3.2). Der Vergleich der beiden Algorithmen erfolgt daher auch nur über die Trefferraten der Richtungsschätzung. Die zum Vergleich herangezogenen Testszenarien sind mit denen aus dem vorangegangenen Abschnitt 5.8 identisch. Es wurde untersucht, inwieweit das GCC-Phat-Verfahren sich bewegende simultane und kreuzende Quellen detektieren bzw. verfolgen kann. Zudem wurden diese Ergebnisse mit dem Resultaten der kombinierten Schätzung verglichen. Die erzielten *Acc*-Trefferraten der Richtungsschätzung für unterschiedliche SNR-Pegel bei realen Impulsantworten und drei verschiedenen Testszenarien sind in Abbildung 5.22 dokumentiert.



(a) 1 sich bewegende Quelle. **(b)** 2 sich aufeinander zubewegende Quellen. **(c)** 2 sich kreuzende Quellen.

Abbildung 5.22: Vergleich der *Acc*-Trefferraten für die Einfallsrichtungsschätzung zwischen GCC-Phat-Verfahren und dem eigenen kombinierten Schätzverfahren.

In Abbildung 5.22a werden die Verfahren für eine sich bewegende Quelle verglichen. Bereits bei nur einem zu detektierenden Sprecher kann sich die kombinierte Schätzung gegen das GCC-Phat-Verfahren durchsetzen. Es ist bei allen SNR Bedingungen überlegen. Es wird vermutet, dass dies vor allem den aufbauenden Erweiterungen der kombinierten Schätzung zuzuschreiben ist. Bei dem Szenario mit sich aufeinander zubewegenden Sprechern, welches für die Abbildung 5.22b als Vorlage diente, fällt das Resultat ebenso deutlich aus. Es ist dabei zu erwähnen, dass die *Acc*-Trefferrate hier kaum unter 20% fallen kann. Die Sprecherpositionen befinden sich im letzten Fünftel des Testszenarios Mittig zum Mikrofon-Array. Wie bereits in Abschnitt 5.2 erwähnt wurde, tendiert die Richtungsschätzung bei

schlechten SNR automatisch zur Ebenenmitte. In diesem Fall scheint die simultane Grundfrequenzschätzung einen erheblichen Beitrag zur Quellenunterscheidung beizutragen. Die Schätzung der dominaten Quelle für erfolgt sicherer als bei dem Beispiel aus Abbildung 5.22a. Es wird vermutet, dass dies durch das Sperr-Filter verursacht wird. Durch zeitweiliges korrektes filtern der zweiten Quelle kann die Trefferate der dominaten Quelle gegenüber der Schätzung nur einer (Abbildung 5.22a) Quelle verbessert werden. Abbildung 5.22c zeigt die sich einstellenden *Acc*-Trefferraten für simultane sich kreuzende Sprecher. In diesem Fall erreichen auch die Trefferraten der kombinierten Schätzung auf Grund der anspruchsvollen Aufgabe nicht die zuvor erzielten Werte bzw. verringern sich deutlich mit sinkendem SNR. Mit schlechteren SNR ist der Algorithmus nicht mehr in der Lage die Sprecher beim Kreuzen zu trennen (vgl. Abschnitt 5.8). Das GCC-Phat-Verfahren erzielt bei diesem Beispiel für eine Quelle Trefferaten im Wertebereich der kombinierten Schätzung. Die zweite Quelle kann, wie bei der Konstellation zuvor jedoch nur sehr schlecht detektiert werden. Es wird wieder davon ausgegangen, dass das Sperr-Filter positiv zur Schätzung der dominanten Quelle beiträgt, die Trefferaten der kombinierten Schätzung für die dominatere Quelle konnten trotzdem nicht errreicht werden. Somit erzielt die kombinierte Schätzung bei den vorliegenden Beispielen in der Summe merklich bessere Resultate als das bekannte GCC-Phat-Verfahren. Dies zeigt das eine kombinierte Grundfrequenschätzung einhergehend mit der Optimierung der Kernalgorithmen durch zusätzliche Erweiterungen einen Mehrwert bei der Einfallrichtungsschätzung von akustischen Signalquellen erwirkt.

Kapitel 6

Zusammenfassung und Ausblick

Zusammenfassung

Im Rahmen dieser Arbeit wurden Algorithmen zur gleichzeitigen, kombinierten Grundfrequenz- und Richtungsschätzung untersucht. Dabei kamen im Kern zwei verschiedene Verfahren zum Einsatz. Diese beruhen zum einen auf zeitlichen Signalinformationen (KKF) und zum anderen auf spektralen Repräsentationen (KLDS) der Eingangssignale zur Ermittlung der Quellencharakteristiken. Zu diesen beiden Kernfunktionen wurden zusätzliche Erweiterungen entwickelt und analysiert, die zu einer Verbesserung der Gesamtleistung beitragen. Dabei handelte es sich um eine in der Literatur beschriebene Cepstrums-Gewichtung, Filterbankvorverarbeitung und einfache Mehrkanalerweiterung. Zusätzlich kamen selbst entwickelte Erweiterungen der GCC-Phat-Gewichtung und MCCC-Mehrkanalkombination hinzu. Das Resultat dieser Verfahren diente als Pseudo-Likelihood-Funktion für ein Partikel-Filter. Durch den Einsatz des Partikel-Filters wurde es ermöglicht die Einfallsrichtung und Grundfrequenz einer Quelle zu einem Zeitpunkt automatisch zu bestimmen und über einen längeren Zeitverlauf einem Sprecher zuzuordnen. Das Partikel-Filter wurde um ein Sperr-Filter erweitert um ebenso simultane Mehrsprechersituationen handhaben zu können.

Als Testszenario wurde ein Mikrofon-Line-Array mit sechs Kanälen verwendet, in dessen Umgebung neun verschiedene Sprecherpositionen mit Hilfe von simulierten oder realen Impulsantworten als Quellen dienten. Inhalt der Evaluation war es zunächst herauszufinden welche Kombination aller beschriebenen Verfahren und Parameter zur besten Leistungsfähigkeit führte. Dabei zeigte es sich, dass die spektrale Schätzung im Kontext des erarbeiteten Systems der Version im Zeitbereich überlegen war. Es konnte zudem gezeigt werden, dass die unterstützenden Erweiterungen einen positiven Einfluss auf die Gesamtleistung des Systems haben. Die resultierende Kombination aller Verfahren und Paramertergrößen, die die besten Ergebnisse lieferte, ist:

- KLDS-Kernfunktion mit $T_2\{\cdot\}$ Phasentransformation
- MCCC-Mehrkanalkombination,
- GCC-Phat-Gewichtung,
- Cepstrum-Gewichtung,

- Filterbankvorverarbeitung $F = 64$,
- Partikelfilter mit 50 Elementen,
- Langevin-Modell mit $\beta_{x,y} = 10\,\mathrm{s}^{-1}$, $v_{x,y} = 1\,\mathrm{m/s}$, $\beta_{f_0} = 5\,\mathrm{s}^{-1}$

Mit dem erarbeiteten System wurden anschließend erweiterte Tests unter verschiedenen Störgräuschpegeln und Nachhallzeiten durchgeführt. Außerdem wurde untersucht, in wie weit die kombinierte Schätzung in der Lage ist, simultane sich bewegende und kreuzende Sprecher zu detektieren und über die Zeit korrekt zu verfolgen. Es lässt sich festhalten, dass bei der Verwendung der realen Impulsantworten mit ca. 550 ms Nachhall bis zu einem SNR von 15 dB gute bis sehr gute Trefferraten auch für sich bewegende und kreuzende Sprecher erzielt werden konnten. Zu guter Letzt wurde das aufgestellte System mit dem wohlbekannten GCC-Phat-Verfahren vergleichen. Dabei dienten die Resultate der Schätzverfahren jeweils als Eingangssignale für das Partikel-Filter. Als Resümee lässt sich feststellen, dass die kombinierte Schätzung besonders bei zwei simultanen und sich kreuzenden Sprechern dem einfachen GCC-Phat-Verfahren deutlich überlegen ist.

Ausblick

Die sich bei Freisprechanlagen und Überwachungssystemen durch automatische Sprechererkennung, -bestimmung und -verfolgung ergebenden Anwendungsszenarien rechtfertigen auch in naher Zukunft eine Erforschung von Lokalisationsalgorithmen. Dabei sollte das Hauptaugenmerk in der Verbesserung der simultanen Sprechererkennung (Anzahl der Sprecher) und einer Steigerung der Robustheit gegenüber Störeinflüssen liegen. Auch bei dem untersuchten Algorithmus gibt es noch Verbesserungspotential, welches im Rahmen dieser Arbeit auf Grund des Zeitrahmens nicht näher untersucht werden konnte. Dabei ist die Optimierung des Sperr-Filters zur Mehrsprecherdiskrimination zu nennen. Dies gilt besonders für Szenarien bei denen sich die Sprecher kreuzen können. Es sollten zudem weitere Partikel-Filter-Verfahren untersucht werden, die unter anderem die Anzahl und Lebensdauer zu untersuchender Sprecher automatisch erkennbar machen. Der Ansatz von zum Beispiel [SVL05] scheint in diesem Zusammenhang Potential zu bieten. Eine weitere in Rahmen dieser Arbeit nicht weiter untersuchte Optimierung stellt die Array-Geometrie dar. Außerdem ist die kombinierte Schätzung auf dreidimensionale Lokalisation zu erweitern. Auch die Kombination mit bildverarbeitenden Verfahren zur z.B. Gesichtserkennung [KE06] verdient einer weiteren Betrachtung. Es wäre auch interessant zu untersuchen, inwieweit durch die Grundfrequenzschätzung eine Unterdrückung unerwünschter und räumlich kohärenter Signale, die keine harmonische Struktur aufweisen (z.B Türschlagen oder Klopfen), möglich ist. Die sich unmittelbar an diese Arbeit anschließenden Aufgaben würden jedoch zunächst eine genauere Untersuchung der MCCC-Mehrkanalberechnung (vgl. Abschnitt 3.4) beeinhalten. Es bedarf noch weiteren Aufwand, um die dem MCCC-Verfahren inherenten Mechanismen tiefgreifender

zu ergründen. Ebenso sollte untersucht werden, ob die Gammatonefilterbank (vgl. Abschnitt 3.6) nicht auch durch eine einfachere Aufteilungen des Spektrums in mehrere FFT-Bandfilter ersetzt werden kann, oder aber deren psychoakustische Motivation weitgreifendere Auswirkungen hat. Unter dem Hintergrund der breiten Parameterwahl bei den Phasentransformationen (vgl. Abschnitt 3.3 und 5.5), lohnt sich auch dort ein weiterer Blick auf deren Bedeutung für das Gesamtsystem der kombinierten Schätzung und die Frage ob nicht noch geeignetere Funktionen gefunden werden können.

Danksagung

An dieser Stelle möchte ich mich ganz herzlich bei allen bedanken, die es mir ermöglicht haben diese Arbeit zu verfassen. Da sind zuerst Prof. Simon Doclo für die freundliche Betreuung als Erstprüfer sowie Stefan Goetze als fachlicher Betreuer bei der Projektgruppe HSA des Fraunhofer Instituts zu nennen. Ihnen danke ich für die tolle Betreuung, Hilfestellung, fachliche Eingabe, Motivation aber auch konstruktive Kritik. Prof. Jörg Bitzer danke ich für die freundliche Übernahme des Zweitprüferamtes sowie den hilfreichen Hinweisen aus unvoreingenommener Sicht. Besonders möchte ich mich bei Henning Schepker für die zahlreichen, weiterführenden Gespräche und Diskussion bis in die späten Abendstunden bedanken, die des öfteren Sachverhalte neu erblicken ließen und Problemstellungen gelöst haben. Ein Dankeschöne geht natürlich auch an die gesamte Belegschaft der Fraunhofer Projektgruppe HSA, welche ein wirklich gutes Klima zur Bewältigung solch einer Arbeit mit stets offenen Ohren geschaffen haben. Ein außerordentliches Dankeschön geht an meine Eltern, ohne deren Einverständis und Unterstützung ich das Masterstudium nicht hätte bewältigen können. Ein schöner, interessanter aber auch anstrengender und nicht zu letzt prägender Abschnitt meines Lebens findet mit dieser Arbeit ein wundervolles Ende.

Eidesstattliche Versicherung

Hiermit versichere ich, dass ich diese Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Außerdem versichere ich, dass ich die allgemeinen Prinzipien wissenschafltlicher Arbeit und Veröffentlichung, wie sie in den Leitlinien guter wissenschaftlicher Praxis der Carl von Ossietzky Universität Oldenburg festgelgt sind, befolgt habe.

Oldenburg,

. . . . . .

Abkürzungsverzeichnis

SNR	Signal-Rausch-Verhältnis (engl. Signal to Noise Ratio).
AKF	Autokorrelationsfunktion.
ALDS	Autoleistungsdichtespektrum.
bzw.	beziehungsweise.
ca.	circa.
dB	Hilfsmaßeinheit Dezibel.
deut.	deutsch.
DFT	Discrete Fourier Transformation.
DoA	Direction of Arrival.
engl.	englisch.
ERB	Equivalent Rectangular Bandwidth (deut. Äquivalentrechteck-Bandbreite).
FFT	Fast Fourier Transformation.
GCC	Generalized Cross-Correlation.
Hann	Hanning Fenster.
IDFT	Inverse Discrete Fourier Transformation.
IFFT	Inverse Fast Fourier Transformation.
KKF	Kreuzkorrelationsfunktion.
KLDS	Kreuzleistungsdichtespektrum.
LDS	Leistungsdichtespektrum.
MCCC	Multichannel Cross-Correlation.
MSC	Magnitude Squared Coherence (deut. Betragsquadrat der Kohärenzfunktion).
Phat	Phasen Transformation (engl. Phat Transform).

PoPi	Position Pitch Algorithm (deut. Positions Grundfrequenz Algorithmus).
RIR	Raumimpulsantwort (engl. Room Impulse Response).
TDD	Time Delay Difference.
TDE	Time Delay Estimation.
TDoA	Time Difference of Arrival Estimation.
ToA	Time of Arrival Estimation.
vgl.	vergleiche.
z.B.	zum Beispiel.

Formelzeichenverzeichnis

Notationen

$\breve{\cdot}$	Sperrfilterung.
$\hat{\cdot}$	Schätzgröße.
$\cdot^{+}$	Halbwellengleichgerichtet.
$\cdot^{*}$	Konjugiert komplex.
$\cdot_{\mathbf{cep}}$	Cepstrum gewichtet.
$\cdot_{\mathbf{phat}}$	Phat (Phase Transform) gewichtet.
$\cdot_{\mathbf{freq}}$	Frequenzbereich.
$\cdot_{\mathbf{g}}$	Gammatonefilter.
$\cdot_{\mathbf{hyb}}$	Hybird.
$\cdot_{i}$	Laufindex.
$\cdot_{\ell}$	Blockindex.
$\cdot_{l}$	Laufindex.
$\cdot_{\max}$	Maximalwert.
$\cdot_{mov}$	Fließende Maximumssuche.
$\cdot_{p}$	Laufindex.
$\cdot_{\mathbf{real}}$	Realer Wert.
$\cdot_{\mathbf{zeit}}$	Zeitbereich.
$\cdot_{\mathbf{zp}}$	Zeropadding.

Operatoren

$\|\cdot\|_2$	Euklidischer Abstand.
$\angle$	Phase des komplexen Signals.
$*$	Faltung.
$\mathbf{det}(\cdot)$	Determinante der Matrix.

$\mathbf{DFT}\{\cdot\}$	Zeitdiskrete Fourier-Transformationsfunktion.
$\mathbf{E}\{\cdot\}$	Erwartungswert.
$\in$	Element einer Menge.
$\mathbf{IDFT}\{\cdot\}$	Inverse zeitdiskrete Fourier-Transformationsfunktion.
$\mathbf{T}\{\cdot\}$	Transformatorfunktion für die Phasengewichtung.

Symbole

$\gamma[n]$	Kohärenzfunktion.
λ	Wellenlänge.
υ	Faktor des Zeropadding.
a	Dämpfungsfaktor.
Acc	Relative Trefferrate.
α	Rekursiver Glättungsfaktor.
b	Diskreter Quefrenzindex, allgemeiner Laufindex.
β	Kontrollparameter.
$br(\cdot)$	Notchbreite.
c	Schallgeschwindigkeit mit $343m/s$.
d	Abstand zwischen zwei Punkten.
$\exp(\cdot)$	Exponentialfunktion $e^{(\cdot)}$.
F	Anzahl der Filter.
$\mathbf{F}(\cdot,\cdot)$	Likelihood Funktion.
f	Frequenz.
f_0	Grundfrequenz der Stimme (Pitch).
f_s	Abtastfrequenz.
GCC	Generalized Cross-Correlation (GCC).
$h[k]$	Impulsantwort.
k	Diskreter Zeitindex, allgemeiner Laufindex.
κ	Zeitdiskrete Verschiebung, Laufindex.
L	Anzahl von Partikel.
l	Kartesische Koordinaten eines Punktes $l = [x, y, z]$.
M	Anzahl der Mikrofone.

m	Mikrofon.
N'	Länge des Signalvektors mit Zeropadding.
N	Blocklänge.
$\mathbb{N}$	Bereich der natürlichen Zahlen $0, 1, 2, \ldots$.
n	Diskreter Frequenzindex, allgemeiner Laufindex.
n_0	Frequenzbin.
$\nu_0(f_0)$	Von f_0 abhängiger Abstand in Samples.
$\Omega(\cdot)$	Normierte Kreisfrequenz.
P	Anzahl der betrachteten Indizes.
$\check{\mathbf{P}}(\cdot)$	Vereinfachte MCCC Matrix.
$\mathbf{P}(\cdot)$	MCCC Matrix.
$p(\cdot\|\cdot)$	Bedingte Wahscheinlichkeit.
φ	Azimutwinkel.
$\phi[n]$	Leistungsdichtespektrum (LDS).
$\Psi(\cdot)$	Notch-Funktion.
ψ	Phase.
Q	Anzahl Quellen.
$q[k]$	Rauschsignal.
$\mathcal{R}$	Rechenoperationen.
$\mathbb{R}$	Bereich der reellen Zahlen.
$\delta[\cdot]$	Deltapulsfolge.
$r[\cdot]$	Korrelation.
$\mathbf{r}$	Gewichtungsvariable.
$\rho(\cdot)$	DoA und Grundfrequenz Schätzebene mittels KKF ermittelt.
$\tilde{\rho}(\cdot)$	DoA und Grundfrequenz Schätzebene mittels LDS ermittelt.
$\check{\rho}(\cdot)$	DoA und Grundfrequenz Schätzebene mittels MCCC ermittelt.
$\rho'(\cdot)$	Auf Maximalpegel normierte DoA und Grundfrequenz Schätzebene.
$s[k]$	Nutzsignal.

$\mathbf{s}$	Partikel.
T	Periodendauer.
t	Kontinuierliche Zeitvariable.
τ_{60}	Nachhallzeit.
τ	Laufzeitunterschied in Sekunden.
V	Blockvorschub.
t_s	Glättungszeit.
u	Gleichverteilte Zufallsvariable.
v	Geschwindigkeit in m/s.
$w[\cdot]$	Cepstrumgewichtung.
$\mathbf{w}$	Partikelgewichtung.
$x[\cdot]$	Mikrofonsignal.
$z[\cdot]$	GCC Filter.

Literaturverzeichnis

[AMGC02] Arulampalam, M., Maskell, S., Gordon, N. und Clapp, T.: *A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking*. Signal Processing, IEEE Transactions on, 50(2):174–188, Feb. 2002. (Zitiert auf den Seiten 52 und 56.)

[BCH04] Benesty, J., Chen, J. und Huang, Y.: *Time-delay estimation via linear interpolation and cross correlation*. Speech and Audio Processing, IEEE Transactions on, 12(5):509–519, Sep. 2004. (Zitiert auf Seite 31.)

[Bit01] Bitzer, J.: *Mehrkanalige Geräuschunterdrückungssysteme - eine vergleichende Analyse*. Doktorarbeit, Universität Bremen, Fachbereich 1 (Physik/Elektrotechnik), 2001. (Zitiert auf Seite 64.)

[Bla74] Blauert, J.: *Räumliches Hören*. Hirzel Verlag, Stuttgart, 1974. (Zitiert auf Seite 15.)

[BOS08] Brutti, A., Omologo, M. und Svaizer, P.: *Localization of multiple speakers based on a two step acoustic map analysis*. In: *Acoustics, Speech and Signal Processing, ICASSP. IEEE International Conference on*, Seiten 4349–4352, April 2008. (Zitiert auf Seite 59.)

[BSH08] Benesty, J., Sondhi, M. und Huang, Y.: *Time delay estimation and source localization*. Springer handbook of speech processing, Seiten 1043–1064, 2008. (Zitiert auf Seite 13.)

[BSM06] Bronstein, I., Semendjajew, K. und Musiol, G.: *Taschenbuch der Mathematik*. Harri Deutsch Verlag, 2006. (Zitiert auf Seite 4.)

[CBH03] Chen, J., Benesty, J. und Huang, Y.: *Robust time delay estimation exploiting redundancy among multiple microphones*. Speech and Audio Processing, IEEE Transactions on, 11(6):549–557, Nov. 2003. (Zitiert auf den Seiten 31, 40 und 41.)

[CBH06] Chen, J., Benesty, J. und Huang, Y. A.: *Time delay estimation in room acoustic environments: an overview*. EURASIP Journal on Applied Signal Processing, ID 26503:1–19, 2006. (Zitiert auf den Seiten 13, 14, 15, 26 und 40.)

[CHB05] Chen, J., Huang, Y. und Benesty, J.: *Time delay estimation via multichannel cross-correlation [audio signal processing applications]*. Band 3, Seiten iii/49–iii/52, Mar. 2005. (Zitiert auf Seite 40.)

[dCK02] Cheveigné, A. de und Kawahara, H.: *YIN, a fundamental frequency estimator for speech and music*. The Journal of the Acoustical Society of America, 111(4):1917–1930, 2002. (Zitiert auf Seite 8.)

[Ell09] Ellis, D. P. W.: *Gammatone-like spectrograms*. web resource, 2009. Last updated: Date: 2009/07/07 14:14:11. (Zitiert auf Seite 46.)

[GKM06] Goetze, S., Kammeyer, K.-D. und Mildner, V.: *A psychoacoustic noise reduction approach for stereo hands-free systems*. In: *Audio Engineering Society Convention 120*, May 2006. (Zitiert auf Seite 64.)

[Goe10] Goetze, S.: *Lecture skript: speech and audio signal processing*. Technischer Bericht, University of Bremen (Dept. of Communication Engineering), 2010. (Zitiert auf Seite 2.)

[Hab10] Habets, E.: *Room impulse response generator*, June 2010. (Zitiert auf den Seiten 15, 16 und 63.)

[HKO08] Habib, T., Kepesi, M. und Ottowitz, L.: *Experimental evaluation of the joint position-pitch estimation (POPI) algorithm in noisy environments*. In: *Sensor Array and Multichannel Signal Processing Workshop, SAM, 5th IEEE*, Seiten 369–372, July 2008. (Zitiert auf den Seiten 31, 42 und 65.)

[HOK08] Habib, T., Ottowitz, L. und Képesi, M.: *Experimental evaluation of multi-band position-pitch estimation (M-PoPi) algorithm for multi-speaker localization*. In: *Interspeech*, Seiten 1317–1320, 2008. (Zitiert auf den Seiten 29 und 49.)

[Hol07] Holube, I.: *Vorlesungsskript Medizinische Akustik*. Jadehochschule Oldenburg, 2007. (Zitiert auf den Seiten 5 und 6.)

[HR10] Habib, T. und Romsdorfer, H.: *Comparison of SRP-Phat and multiband-Popi algorithms for speaker localization using particle filters*. In: *13th International Conference on Digital Audio Effects, DAFX, Graz, Austria*, Sep. 6-10 2010. (Zitiert auf den Seiten 46, 53, 56 und 66.)

[KC76] Knapp, C. und Carter, G.: *The generalized correlation method for estimation of time delay*. In: *Acoustics, Speech and Signal Processing, IEEE Transactions on*, Band 24, Seiten 320–327, Aug. 1976. (Zitiert auf den Seiten 2 und 27.)

[KE06] Küblbeck, C. und Ernst, A.: *Face detection and tracking in video sequences using the modified census transformation*. Image and Vision Computing, 24(6):564–572, 2006. (Zitiert auf Seite 85.)

[KJ05] Kiencke, U. und Jäkel, H.: *Signale und Systeme*. Oldenbourg Wissenschaftsverlag, 2005. (Zitiert auf den Seiten 17 und 20.)

[KK09] Kammeyer, K.-D. und Kroschel, K.: *Digitale Signalverarbeitung - Filterung und Spektralanalyse mit MATLAB®-Übungen*.

Vieweg+Teubner-Verlag, Wiesbaden, Germany, 7 Auflage, Apr 2009. (Zitiert auf den Seiten 26 und 31.)

[KOH08] Kepesi, M., Ottowitz, L. und Habib, T.: *Joint position-pitch estimation for multiple speaker scenarios.* In: *Hands-Free Speech Communication and Microphone Arrays, HSCMA*, Seiten 85–88, May 2008. (Zitiert auf den Seiten 31, 45, 46 und 47.)

[KPW07] Kepesi, M., Pernkopf, F. und Wohlmayr, M.: *Joint position-pitch tracking for 2-channel audio.* In: *Content-Based Multimedia Indexing, CBMI, International Workshop on*, Seiten 303–306, June 2007. (Zitiert auf den Seiten 29, 31 und 32.)

[Kut07] Kuttruff, H.: *Acoustic: An introduction.* Taylor & Francis, 2007. (Zitiert auf Seite 18.)

[KWH07] Képesi, M., Wohlmayr, M. und Habib, T.: *Pitch-driven position estimation of speakers in multispeaker environments.* In: *3rd Congress of the Alps Adria Acoustics Association*, Graz - Austria, Sep. 2007. (Zitiert auf Seite 34.)

[KWK07] Kepesi, M., Wohlmayer, M. und Kubin, G.: *Patent: joint position-pitch estimation of acoustic sources for their tracking and separation*, June 2007. (Zitiert auf Seite 8.)

[LJ07] Lehmann, E. A. und Johansson, A. M.: *Particle filter with integrated voice activity detection for acoustic source tracking.* EURASIP Journal on Advances in Signal Processing, 2007. Article ID 50870, 11 pages. (Zitiert auf Seite 52.)

[Mey08] Meyer, T.: *Verfahren zur Bestimmung der Einfallsrichtung von Audiosignalen für mehrkanalige Hörgerätealgorithmen (Studienarbeit).* Diplomarbeit, Universität Bremen, Arbeitsbereich Nachrichtentechnik, 2008. (Zitiert auf den Seiten 14 und 40.)

[MM01] Müller, S. und Massarani, P.: *Transfer-function measurement with sweeps.* J. Audio Eng. Soc, 49(6):443–471, 2001. (Zitiert auf Seite 63.)

[Nol67] Noll, A. M.: *Cepstrum pitch determination.* The Journal of the Acoustical Society of America, 41(2):293–309, 1967. (Zitiert auf Seite 11.)

[PK08] Pfister, B. und Kaufmann, T.: *Sprachverarbeitung.* Springer Berlin Heidelberg, 2008. (Zitiert auf den Seiten 5, 6, 7, 8 und 11.)

[PM96] Proakis, J. G. und Manolakis, D. G.: *Digital signal processing: principles, algorithms, and applications.* Upper Saddle River, EUA : Prentice-Hall, 1996. (Zitiert auf Seite 20.)

[Roe07] Roeske, P.: *Sprecherlokalisation in gestörter Umgebung (Diplomarbeit).* Diplomarbeit, Fachhochschule Oldenburg / Ostfriesland / Wil-

helmshaven, Institut für Hörtechnik und Audiologie, 2007. (Zitiert auf Seite 17.)

[SVL05] Särkkä, S., Vehtari, A. und Lampinen, J.: *Rao-blackwellized particle filter for multiple target tracking*. Information Fusion Journal, 8:2007, 2005. (Zitiert auf Seite 85.)

[VHH98] Vary, P., Heute, U. und Hess, W.: *Digitale Sprachsignalverarbeitung*. Teubner, 1998. (Zitiert auf den Seiten 5, 6, 8, 11 und 64.)

[Vol10] Volgenandt, A.: *Implementierung und Evaluation mehrkanaliger Algorithmen zur Lokalisation von mehreren Sprechern in der Konferenzsituation*. Diplomarbeit, Fachhochschule Oldenburg / Ostfriesland / Wilhelmshaven, 2010. (Zitiert auf den Seiten 18 und 59.)

[Wag09] Wagner, J.: *Tutorial for Smart Sensor Integration (SSI)*. University of Augsburg, 2009. (Nicht zitiert.)

[WAJ09] Wagner, J., André, E. und Jung, F.: *Smart sensor integration: A framework for multimodal emotion recognition in real-time*. In: *Affective Computing and Intelligent Interaction (ACII 2009)*, 2009. (Nicht zitiert.)

[WK07] Wohlmayr, M. und Képesi, M.: *Joint Position-Pitch Extraction from Multichannel Audio*. In: *Interspeech*, Seiten 1629–1632, 2007. (Zitiert auf den Seiten 36, 38 und 40.)

[WLW03] Ward, D., Lehmann, E. und Williamson, R.: *Particle filtering algorithms for tracking an acoustic source in a reverberant environment*. Speech and Audio Processing, IEEE Transactions on, 11(6):826–836, nov. 2003. (Zitiert auf den Seiten 52, 54, 66 und 67.)