Directional Audio Coding
Implementierung und experimentelle Bestimmung der optimalen Zeit- und Frequenzauflösung

Seminararbeit aus Algorithmen in Akustik und Computermusik 2

Maximilian Giller
Benjamin Stahl

Betreuung: Dr. Franz Zotter, Dr. Matthias Frank

Graz, 29. Juni 2016

institut für elektronische musik und akustik

Abstract

Diese Arbeit behandelt DirAC (*Directional Audio Coding*), ein Auflösungsverbesserungsverfahren für Surround-Aufnahmen, das auf der Analyse von Richtung und Diffusität der Frequenzkomponenten des Eingangssignals beruht.

Nach einer Erläuterung der Funktionsweise des Algorithmus und der zugrundeliegenden Annahmen aus der Psychoakustik wird das Verfahren in MATLAB implementiert. Verschiedene Parameter bestimmen, wie gut die Ergebnisse sind. Einer dieser Parameter ist die Länge der verwendeten DFT. In einem Hörversuch haben wir für verschiedene Signalarten die ideale DFT-Länge bestimmt. Dabei wurden die Kriterien Lokalisationsschärfe und Klangqualität bewertet.

Es stellte sich heraus, dass die vorliegende Realisierung von DirAC am besten für Sprache geeignet ist. Bei Musik, insbesondere mit deutlichem Nachhall, funktioniert der Algorithmus am schlechtesten.

Abschließend werden Vorschläge zur weiteren Optimierung des Algorithmus vorgestellt.

Inhaltsverzeichnis

1 Einleitung

Die originalgetreue Abbildung eines dreidimensionalen Schallfeldes ist eine besondere Herausforderung für die Aufnahmetechnik. Zur Wiedergabe stehen verschiedene Verfahren zur Verfügung, wie beispielsweise *Surround-Sound 5.1*, *VBAP*, *Wellenfeldsynthese* oder *Ambisonics*.

Es muss allerdings ein gewisser Aufwand betrieben werden, um Aufnahmen mit hoher Lokalisationsschärfe und Audioqualität zu ermöglichen. Entsprechende Mikrofonanordnungen sind oft relativ groß und und haben ein nicht optimales Übertragungsverhalten, wie z.B. das *Eigenmike*, das für Ambisonics höherer Ordnung geeignet ist. Häufig stehen also aufgrund von finanziellen oder technischen Einschränkungen nur Aufnahmen mit einer verhältnismäßig geringen räumlichen Auflösung zur Verfügung, wie beispielsweise Aufnahmen im B-Format 1. Ordnung (Ambisonics 1. Ordnung). Wäre es nicht erstrebenswert, in der Lage zu sein, die Lokalisationsschärfe einer niedrig auflösenden Surround-Aufnahme zu verbessern ohne die Klangqualität zu beeinträchtigen, oder sie gar zu erhöhen?

Bei *Directional Audio Coding* (*DirAC*) handelt es sich um ein Verfahren zur Verbesserung der Richtungsauflösung bei der Wiedergabe von räumlichem Schall über Lautsprecher, das 2007 von Ville Pulkki vorgestellt wurde [1]. Die möglichst naturgetreue Reproduktion realer Klangereignisse ohne Klangverfärbungen bei größtmöglicher Lokalisationsschärfe ist das grundlegende Ziel des Verfahrens. DirAC beruht auf der frequenzabhängigen Richtungs- und Diffusitätsanalyse einer Surround-Aufnahme. Bei der anschließenden Resynthese dienen die ermittelten Informationen zur Kodierung der frequenzabhängigen Richtungskomponenten des Schallfeldes in einem Surround-Wiedergabeformat wie z. B. *VBAP* oder *Ambisonics*, das über eine Lautsprecheranordnung wiedergegeben wird.

Die Lokalisierbarkeit soll dadurch verbessert werden, dass weniger Lautsprecher maßgeblich und mit kohärentem Signal zur Repräsentation des Direktschalls beitragen. Dabei soll auch eine Verbesserung der Klangqualität erreicht werden, welche mit der Verringerung von interferenzbedingten Klangverfärbungen als Folge der geringeren Kohärenz der Lautsprechersignale begründet wird. Damit das Verfahren allerdings selbst möglichst wenige Artefakte und Verfärbungen produziert, muss die Parametrierung des Algorithmus mit einiger Sorgfalt erfolgen.

In dieser Arbeit wird eine grundlegende Implementierung von DirAC beschrieben und getestet. Das folgende Kapitel (Kap. 2) beschäftigt sich mit den theoretischen Grundlagen, der Funktionsweise und der konkreten Realisierung des Algorithmus. In Kapitel 3 wird ein Hörversuch durchgeführt, um optimale Einstellungsparameter zu finden. Eine Zusammenfassung und mögliche zukünftige Ansätze zur Verbesserung des Algorithmus finden sich in Kapitel 4.

2 Directional Audio Coding

2.1 Zugrundeliegende Annahmen aus der Psychoakustik

DirAC beruht auf verschiedenen Erkenntnissen und Annahmen bezüglich der Wahrnehmung von räumlichem Schall, auf die im Folgenden kurz eingegangen werden soll.

Die Richtungsbestimmung des menschlichen Gehörs stützt sich auf interaurale Laufzeitdifferenzen (ITD, *Interaural Time Differences*), interaurale Pegeldifferenzen (*ILD, Interaural Level Differences*) und Filterung durch die (monauralen) Übertragungsfuntionen, die durch Außenohr, Kopf und Torso gegeben ist (Außenohrübertragungsfunktion oder *engl.* HRTF, *Head-related Transfer Function*).

Die Verarbeitung im menschlichen Gehör findet aufgeteilt in 24 Frequenzgruppen statt. Schallereignisse, die zusammen in eine Frequenzgruppe fallen, werden unter anderem bei der Lokalisierung gemeinsam verarbeitet. Innerhalb einer Frequenzgruppe ist die Anzahl der wahrnehmbaren Schallquellen begrenzt – im Gegensatz zum Sehsinn, der es uns problemlos ermöglicht, mehrere Objekte mit der gleichen Farbe wahrzunehmen, erlaubt uns das Gehör in der Regel nur, eine einzige Quelle innerhalb einer Frequenzgruppe wahrzunehmen und einer Richtung zuzuordnen [2, S. 1].

Ein wesentlicher Faktor, der sich auf die Lokalisierbarkeit von Schallereignissen auswirkt, ist die *interaurale Kohärenz*, also die Ähnlichkeit der Signale am linken und rechten Ohr. Wenn beispielsweise jedem Ohr über Kopfhörer dasselbe Signal zugeführt wird, entsteht eine Phantomschallquelle in der Kopfmitte: Hier ist der interaurale Kohärenzgrad, der Werte zwischen 0 und 1 annimmt, maximal, da die beiden gehörten Signale identisch sind. Wenn stattdessen zwei unterschiedliche Signale (mit geringerer Kohärenz) verwendet werden, werden zwei getrennte Hörereignisse in der Nähe der Lautsprecher auf der lateralen Achse wahrgenommen [3, S. 105 f.].

Beim Hören in realen Schallfeldern werden die Ohrsignale durch Reflexionen dekorreliert. Jedoch kann im freien Schallfeld der interaurale Kohärenzgrad aufgrund des Übersprechens zwischen den Ohren nie ganz verschwinden. Sind die Ohrsignale teilweise kohärent, führen sie zu räumlich ausgedehnten, diffusen Hörereignissen [3, S. 106].

Bei DirAC wird nun vorausgesetzt, dass das menschliche Gehör zu jedem Zeitpunkt jeder Frequenzgruppe aufgrund von ILD, ITD und HRTFs nur eine Richtung zuordnet. Dies gilt auch für die interaurale Kohärenz. Darüberhinaus wird angenommen, dass diese *cues* korrekt wahrgenommen werden, wenn Richtung und Diffusität des Schallfeldes korrekt gemessen und reproduziert werden [2, S. 1]. Analyse und Synthese erfolgen hier wieder zu jedem Zeitpunkt einmal für jedes Frequenzband (nicht zwingend identisch mit den Frequenzgruppen).

2.2 Funktionsweise des Algorithmus

Hier soll die generelle Funktionsweise des Algorithmus erklärt werden. Dabei wird auf einige die konkrete Implementierung betreffende Details verzichtet, auf die später eingegangen wird. Der Algorithmus wird mit Bezug auf [1] erklärt. Bei der von uns gewählten Implementierung gibt es einige Abweichungen, die ebenfalls später erläutert werden. In Abbildung 1 ist das Verfahren graphisch dargestellt.

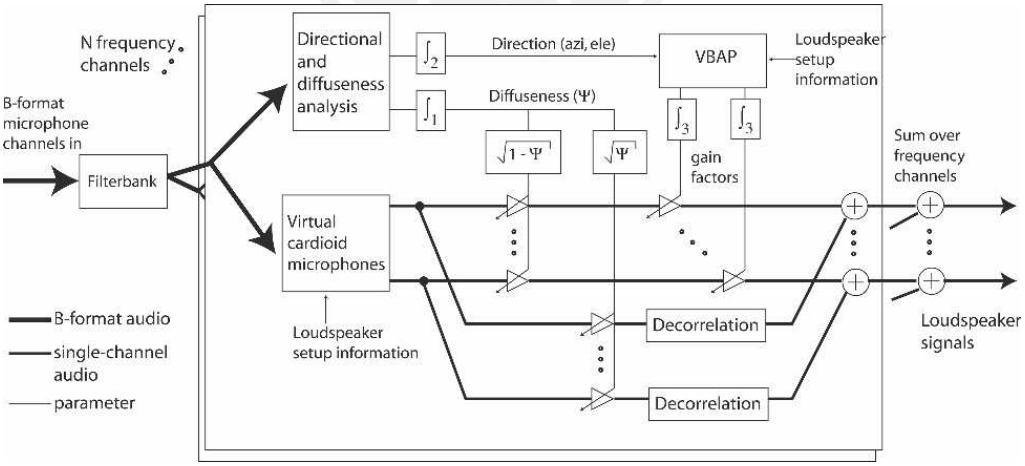
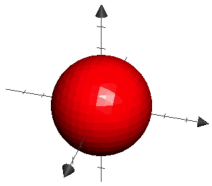


Abbildung 1 – Signalflussdiagramm zur Funktionsweise von DirAC, [1, S. 505].

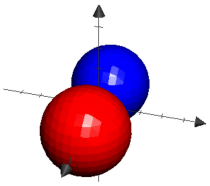
Als Eingangssignal wird ein Signal im vierkanaligen *B-Format* verwendet. Eine Mikrofonanordnung, die für Aufnahmen im B-Format geeignet ist, besteht theoretisch aus einem Druckempfänger mit kugelförmiger Richtcharakteristik in der Mitte der Anordnung, welcher das Signal W liefert, und drei Druckdifferenzempfängern (Achtförmige Richtcharakteristik), die schnelleproportionale Signale X, Y, Z in alle drei Raumrichtungen liefern. Die Richtcharakteristiken zu den verschiedenen Signalen sind in Abb. 2 dargestellt. In der Praxis wird häufig ein sogenanntes *Soundfield-Mikrofon* verwendet. Aus den Signalen der vier tetraederförmig angeordneten Mikrofonkapseln können W, X, Y und Z berechnet werden.

Am Anfang des Verfahrens steht eine Frequenzanalyse, die mittels einer Kurzzeittransformation (STFT) oder einer Filterbank auf die vier Eingangssignale angewendet wird. Der zweite Schritt ist die Bestimmung von Richtung und Diffusität im Zeit-Frequenz-Bereich. Die Bestimmung erfolgt also in bestimmten Zeitabständen für jedes Frequenzband und liefert jeweils eine Funktion für den Azimutwinkel φ, den Elevationswinkel ϑ und die Diffusität ψ.

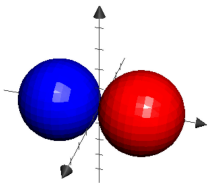
Für die weitere Verarbeitung wird nur noch das omnidirektionale W-Signal verwendet. Dieses wird in eine direkte und eine diffuse Komponente aufgeteilt. Die Diffuskomponente ergibt sich für jede Frequenz k und jeden Signalblock n durch Multiplikation des W-Signals mit der Wurzel der Diffusität $\sqrt{\psi}$, die Werte zwischen 0 und 1 annimmt. Die Direktkomponente ergibt sich analog durch Multiplikation mit $\sqrt{1-\psi}$; durch das Ziehen der Wurzel wird gewährleistet, dass die Energie der beiden Komponenten in Summe nicht



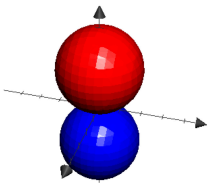
(a) W-Komponente



(b) X-Komponente



(c) Y-Komponente



(d) Z-Komponente

Abbildung 2 – Veranschaulichung der Richtcharakteristik der vier Komponenten W, X, Y, Z des B-Formats.

verändert wird.

Mittels der bestimmten Richtungswerte werden die N zeit- und frequenzabhängigen VBAP-Verstärkungsfaktoren für die gewählte Lautsprecheraufstellung berechnet. Die Direktkomponente wird den N Lautsprechern zugeführt und jeweils mit dem entsprechenden Faktor gewichtet. Zu jedem dieser Lautsprechersignale wird ein Diffussignal addiert, das vorher, beispielsweise durch Faltung mit Zufallsfolgen, dekorreliert wurde. Jeder Lautsprecher enthält also die Summe aus einem Direktanteil mit Richtung und Diffusität entsprechend zeitabhängig gewichteten Frequenzen und einem Diffusanteil, der für jeden Lautsprecher einzeln dekorreliert wurde.

Die so erzeugten Lautsprechersignale werden nun blockweise in den Zeitbereich zurücktransformiert und mittels Overlap-Add zusammengesetzt.

Da die Richtungs- und Diffusitätswerte sich bei entsprechenden Signalen sehr schnell verändern können, entstehen Artefakte. Um dem entgegenzuwirken, müssen die Richtungs- und die Diffusitätsfunktionen mit einem gleitenden Mittelwertfilter geglättet werden.

2.3 Implementierung

In diesem Kapitel wird im Detail beschrieben, wie der DirAC-Algorithmus von uns realisiert wurde. Die Implementierung erfolgte in Form eines MATLAB-Skripts.

Die B-Format-Kanäle $w[n], x[n], y[n]$ und $z[n]$ der zu verarbeitenden Aufnahme werden jeweils zunächst mit 50 % Overlap blockweise gefenstert und in den Frequenzbereich transformiert (STFT). Um theoretisch die ideale Rekonstruktion bei der inversen STFT zu ermöglichen, muss eine Fensterform verwendet werden, die sich bei 50 % Überlappung auf 1 ergänzt. Das Hann-Fenster erfüllt diese Voraussetzung.

Die folgenden Operationen werden, sofern nicht anders angegeben, blockweise im Frequenzbereich durchgeführt.

Der Schnellevektor $\mathbf{V}_m[k] = [X_m[k] \quad Y_m[k] \quad Z_m[k]]$ wird nun verwendet, um für jede Frequenz Schalleinfallsrichtung und Diffusität zu bestimmen, während die Audiodaten, die mithilfe der zuvor berechneten Richtungs- und Diffusitätswerte verarbeitet werden, nur vom omnidirektionalen W-Signal stammen. In obiger Gleichung steht k für den Frequenzindex und m ist der Index des aktuell verarbeiteten Blockes.

2.3.1 Bestimmung von Richtung und Diffusität

Die Schalleinfallsrichtung ist bereits durch $-\mathbf{V}$ gegeben. Da jedoch in der Schallschnelle auch Blindanteile enthalten sind, wird die Richtung

$$\mathbf{D}_m[k] = \begin{bmatrix} d_{1,m}[k] \\ d_{2,m}[k] \\ d_{3,m}[k] \end{bmatrix} = -\mathbf{I}_m[k] = -\mathfrak{Re}\Big\{\operatorname{conj} P_m[k] \cdot \mathbf{V}_m[k]\Big\} \tag{1}$$

in kartesischen Koordinaten aus dem umgekehrten Schallintensitätsvektor bestimmt. In obiger Gleichung bezeichnet P das mit $\sqrt{2}$ gewichtete W-Signal im Frequenzbereich.

Die Richtung in sphärischen Koordinaten ergibt sich wie folgt:

$$\varphi = \operatorname{atan2}\left(\bar{d}_2, \bar{d}_1\right) \quad \ldots \text{Azimutwinkel} \tag{2}$$

$$\vartheta = \operatorname{acos}\frac{\bar{d}_3}{\sqrt{\bar{d}_1^2+\bar{d}_2^2+\bar{d}_3^2}} \quad \ldots \text{Elevationswinkel.} \tag{3}$$

In obiger Gleichung sind $\bar{d}_i$ die Komponenten des gemittelten Richtungsvektors in kartesischen Koordinaten, der mithilfe eines zentrierten gleitenden Mittelwertfilters aus einem Speicher der vorhergehenden und zukünftigen Intensitätswerten implementiert wurde. Die Anzahl der Blöcke, über die gemittelt wird, ist einer der bestimmenden Parameter bezüglich der erzielten Resultate und der Berechnungsdauer.

Die Diffusität des Schallfeldes am Messort

$$\psi = 1 - \frac{||\mathbb{E}\{\mathbf{I}\}||}{c\mathbb{E}\{E\}} \quad \text{mit} \quad E = \frac{1}{2}\rho_0\left(\frac{p^2}{Z_0^2} + v^2\right) \tag{4}$$

kann berechnet werden, indem man den Betrag des Intensitätsvektors mit der Energie vergleicht. In der folgenden Gleichung ist c die Ausbreitungsgeschwindigkeit von Schall, ρ_0 die Dichte von Luft und Z_0 die Schallkennimpedanz.

Es zeigte sich jedoch bei der praktischen Anwendung von Gl. 4, dass der Diffusanteil deutlich überbewertet wurde ($\psi \approx 1$). Sinnvollere Ergebnisse wurden mit Gl. 5 erzielt, die auf dem Verhältnis von der Länge des gemittelten Intensitätsvektors zu seiner gemittelten Länge basiert [4, S. 287]:

$$\psi = \sqrt{1 - \frac{||\mathbb{E}\{\mathbf{I}\}||}{\mathbb{E}\{||\mathbf{I}||\}}}. \tag{5}$$

2.3.2 Mittelung

Sowohl bei der Richtungs- als auch bei der Diffusitätsbestimmung muss gemittelt werden. Bei der Richtungsbestimmung dient die Mittelung des Richtungsvektors dazu, zu schnelle Richtungsschwankungen, die Artefakte verursachen würden, zu vermeiden. Bei der Diffusitätsberechnung kann es ebenfalls notwendig sein, zu schnelle Schwankungen zu vermeiden. Daneben ist jedoch unabhängig von der Vermeidung von Artefakten immer die Bildung des Erwartungswertes des Intensitätsvektors für die Berechnung von ψ notwendig (s. Gl. 5).

Anstatt nun Richtungs- und Diffusitätswerte zu speichern und darüber zu mitteln, wurden Werte des Intensitätsvektors gespeichert und für die Berechnungen verwendet. Um die Verwendung eines zentrierten Filters zu ermöglichen, müssen dabei nicht nur vorhergehende, sondern auch zukünftige Intensitätswerte gespeichert werden.

Realisiert wurde dies mit einem FIFO-Buffer(*first in first out*). Nach einer Initialisierungsphase wird pro Verarbeitungszyklus ein weiteres Element auf dem Stapel abgelegt und das älteste Element auf dem Stapel wird entfernt (vgl. Abb. 3, Block `IV`). Die Gesamtanzahl der Elemente des Buffers ist $M + 1$, und setzt sich zusammen aus $\frac{M}{2}$ alten Elementen ($\mathbf{I}_{m-M/2}, \ldots, \mathbf{I}_{m-1}$), dem aktuellen Element $\mathbf{I}_m$, und $\frac{M}{2}$ zukünfigen Elementen ($\mathbf{I}_{m+1}, \ldots, \mathbf{I}_{m+M/2}$). Wenn M nun das Maximum der gewünschten Filterlängen für Diffusität und Richtung ist, kann der Buffer für zwei verschiedene Filter verwendet werde, von denen das längere Filter über $M + 1$ Werte mittelt, und das kürzere über eine kleinere Anzahl von Werten aus dem selben Buffer.

2.3.3 Vorgehensweise bei der Verarbeitung

Der DirAC-Algorithmus wurde in MATLAB in einer einzigen for-Schleife implementiert. Aus dem Eingangssignal im B-Format werden Blöcke der Länge N mit 50 % Überlappung herausgeschnitten. Abb. 3 zeigt die Verarbeitung zur Berechnung m-ten Signalblocks der Ausgangssignale im Ambisonicsformat (der Direktkomponente $\mathbf{s}_m^{dir}$ und der Diffuskomponente $\mathbf{s}_m^{diff}$). Die folgenden Ausführungen beziehen sich auf diese Abbildung.

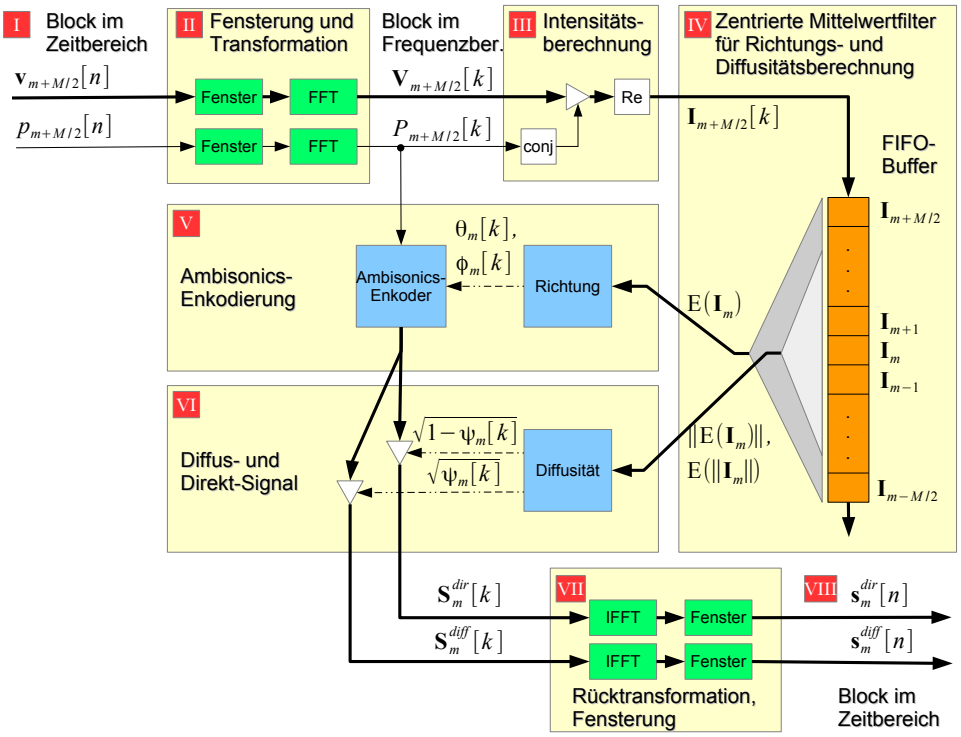


Abbildung 3 – DirAC-Verarbeitung im Frequenzbereich mit MATLAB. Dicke schwarze Pfeile stehen für einen Vektor von Audiosignalen, dünne schwarze Pfeile für ein einzelnes Audiosignal. Die strichlierten Pfeile stehen für Steuersignale.

Da für die Mittelung auch zukünftige Signalwerte benötig werden, liegt zur Berechnung des m-ten Blocks des Ausgangssignals am Eingang des Algorithmus der Signalblock mit Index $m + \frac{M}{2}$ an. Das Eingangssignal besteht aus einem Block des Drucksignals $p_{m+\frac{M}{2}}$ ($p = \sqrt{2}\,w$) und einem Block des Schnellevektors $\mathbf{v}_{m+\frac{M}{2}} = [x_{m+\frac{M}{2}}\ y_{m+\frac{M}{2}}\ z_{m+\frac{M}{2}}]$ (I).

Die Signalblöcke werden zunächst mit einer Fensterfunktion multipliziert und durch eine N-Punkte-FFT den Frequenzbereich transformiert (II). Die Ausgangssignale werden mittels Overlap-Add aus den einzelnen Blöcken am Ausgang (VIII) zusammengesetzt. Es fällt auf, dass in Block VII erneut mit einer Fensterfunktion multipliziert wird. Der Grund dafür ist, dass die Verarbeitung im Frequenzbereich als Filterung betrachtet werden kann, nach der die Energieerhaltung im Zeitbereich bei der Überlagerung der Segmente nicht mehr zwangsläufig gilt und Artefakte erzeugt werden. Dem kann entgegengewirkt werden, indem das Fenster in ein Analyse- und ein Synthesefenster aufgeteilt wird. In unserem Fall ergibt das Produkt von Analyse- und Synthesefenster ein Hann-Fenster, also ein Fenster mit dem die ideale Rekonstruktion gewährleistet ist.

Nachdem also die Signalblöcke den Verarbeitungsblock II durchlaufen haben, liegen Schalldruck und Schallschnellevektor im Frequenzbereich vor. Es folgt in III die Berechnung der Intensität, wie bereits in Gl. 1 beschrieben, durch Berechnung des Realteils des Produkts des komplex konjugierten Schalldrucks mit dem Schnellevektor. Der resultierende frequenzabhängige Intensitätsvektor $\mathbf{I}_{m+\frac{M}{2}}[k]$ wird in dem FIFO-Buffer

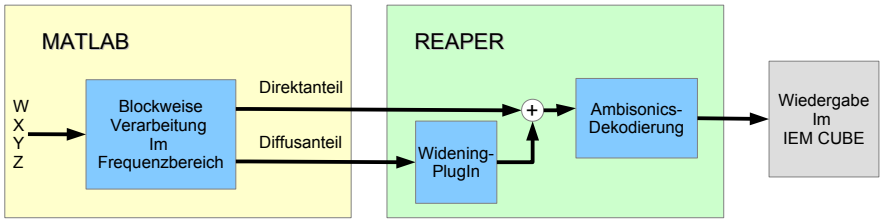


Abbildung 4 – Auf die Verarbeitung in MATLAB folgt die Wiedergabe mit REAPER unter Verwendung eines Widening-Plug-Ins.

(`IV`) gespeichert, der älteste Wert des Buffers wird entfernt.

Nun wird für Diffusitäts- und Intensitätsberechnung jeweils ein Mittelwert aus Elementen des Buffers berechnet. Die Anzahl der dabei von den beiden Berechnungen verwendeten Werte können sich unterscheiden. In `V` erfolgt die Richtungsberechnung wie in Gl. 1 beschrieben. Die resultierenden frequenzabhängigen Azimut- und Elevationswinkel werden dann dafür verwendet, jede Frequenzkomponente des Schalldruckblocks mithilfe eines Ambisonics-Enkoders für die entsprechende Richtung zu kodieren. Wenn Ambisonics mit Ordnung L verwendet wird, erhält man nun für jeden der N Frequenzkanäle $(L+1)^2$ Ambisonics-Komponenten. Das Ergebnis wird in MATLAB also durch eine $N \times (L+1)^2$-Matrix repräsentiert.

Aus diesem Signal im Ambisonics-Format entstehen nun die Direkt- und die Diffuskomponente (`VI`). Je nach berechneter (frequenzabhängiger) Diffusität wird zwischen den beiden Komponenten übergeblendet. Anschließend werden die Blöcke, wie bereits erwähnt, in den Zeitbereich zurücktransformiert erneut gefenstert. Mittels Overlap-Add wird der aktuelle Block an das Ausgangssignal angefügt.

In Abb. 4 ist zu sehen, dass die Ausgangssignale, nachdem sie mit MATLAB vollständig berechnet worden sind, in der DAW (*digital audio workstation*) REAPER weiterverarbeitet werden. Dafür stand die *ambiX Ambisonics plug-in suite* [5] zur Verfügung.

Bis einschließlich Punkt `VIII` in Abb. 3 entspricht die Diffuskomponente lediglich dem Anteil, der nicht der Direktkomponente zugeordnet wurde. Um einen diffusen Klangeindruck zu erzeugen, muss eine weitere Verarbeitung stattfinden. Dazu haben wir das *ambix_widening*-Plug-In verwendet, das durch zeitvariante Rotation der Frequenzkomponenten im Ambisonics-Schallfeld einen diffusen Schalleindruck erzeugt [5, S. 2]. Bei der Wiedergabe werden Direkt- und Diffuskomponente addiert. Die Wiedergabe fand im IEM-CUBE unter Verwendung der dort installierten Hemisphäre aus 24 Lautsprechern statt. Praktischerweise war es zusätzlich möglich, mit dem *ambix_binaural*-Plug-In mittels der binauralen Lautsprecherimpulsantworten aus dem CUBE die Ergebnisse bereits über Kopfhörer anzuhören, was sich als äußerst nützliches Werkzeug zur Evaluation des Algorithmus herausstellte.

3 Hörversuch

In einem Hörversuch wurde nun die im vorigen Abschnitt beschriebene Implementierung für unterschiedliche akustische Situationen getestet. Hierbei wurde für ein Musik- und ein Sprachbeispiel das DirAC-Verfahren mit unterschiedlichen Fensterlängen durchgeführt. Dafür wurden mithilfe einer raumakustischen Simulation B-Format-Signale erstellt. Für jedes Szenario (Sprache/Musik) wurden zwei Beispiele mit unterschiedlichem Nachhallanteil erstellt. Die auf diese Weise generierten Audiobeispiele wurden von den Versuchspersonen in Hinblick auf Lokalisierbarkeit und Klangqualität bewertet.

3.1 Ziele und Hypothesen

Ziel des Hörversuchs war die Beantwortung folgender Forschungsfragen.

- Welche Fensterlängen sind beim DirAC-Verfahren in unterschiedlichen akustischen Szenarien optimal im Hinblick auf Audioqualität und Lokalisationsschärfe?
- Bis hin zu welcher Komplexität des akustischen Szenarios existiert eine Fensterlänge, mit der das DirAC-Verfahren ohne Einbußen bei der Audioqualität (verursacht durch Artefakte und Klangverfärbungen) eingesetzt werden kann?

  Diese Frage kann in folgende Teilfragen aufgetrennt werden:
  - Wie müssen die aufgenommen Klangereignisse geartet sein? Kann das Verfahren ebenso auf Musik wie auf Sprache angewandt werden?
  - Welcher Hallanteil auf der Aufnahme, die als Ausgangsmaterial für das DirAC-Verfahren verwendet wird, ist tolerabel? Der Hallanteil ergibt sich in der Praxis aus dem Aufnahmeabstand und dem Hallradius des Raumes.

Wir vermuteten, dass die Fensterlänge der verwendeten FFT, um mit DirAC ein stabiles Klangbild zu bekommen, für Ausgangsaufnahmen mit einem nicht trivialen Spektrum (Aufnahmen von Sprache und Musik) zumindest zu einer Frequenzauflösung führen sollte, welche nicht dramatisch von der Breite einer Frequenzgruppe des menschlichen Gehörs abweicht. Berücksichtigt man hier, dass diese bei tiefen Frequenzen ca. 100 Hz beträgt, führt dies zu der Hypothese, dass bei einer Abtastrate von 44100 Hz Fensterlängen deutlich unter 441 nicht zum gewünschten Ergebnis führen.

Eine weitere Vermutung, die wir vor dem Experiment anstellten, war, dass die Komplexität, die ein reicheres Spektrum der Ausgangsaufnahme mit sich bringt (wie z.B. bei Musik), sich negativ auf die Verarbeitung mit DirAC auswirkt. Einzelne Schallquellen können nicht mehr räumlich voneinander getrennt werden, weil ihre spektralen Anteile bei unzureichender Frequenzauflösung der FFT in gemeinsame diskrete Frequenzkomponenten fallen.

Wir vermuteten, dass ein großer Diffusanteil auf der Ausgangsaufnahme einen ähnlichen Effekt auf die DirAC-Verarbeitung haben würde, da die Tatsache, dass Spektralanteile,

die nicht im Moment der Aufnahme von den Schallquellen (Sprechern, Instrumenten) abgestrahlt werden, sondern kurze Zeit vorher abgestrahlt wurden, in gefilterter Form als Diffusanteil auf der Aufnahme vorhanden sind, effektiv auch zu einer Verdichtung des verarbeiteten Spektrums führt.

3.2 Versuchsaufbau / Konzept

Wir haben unsere Hypothesen in einem Hörexperiment getestet, in dem die Teilnehmer mehrere Audiobeispiele, darunter das B-Format-Ausgangsmaterial, eine Referenzbeispiel und mehrere DirAC-verarbeitete Beispiele, miteinander vergleichen sollten. In getrennten Sessions sollte einmal das Augenmerk auf Audioqualität bezüglich Klangverfärbungen und Artefakten und einmal auf Lokalisationsschärfe gelegt werden. Sieben Teilnehmer fanden sich für den Hörversuch, darunter die beiden Autoren. Bei allen Teilnehmern handelte es sich um Elektrotechnik-Toningenieur-Studierende. Aufgrund der geringen Teilnehmeranzahl sollte der Hörversuch als Pilotstudie betrachtet werden.

3.2.1 Verwendete Audiobeispiele

Wir verwendeten für unseren Versuch vier unterschiedliche simulierte Ausgangsaufnahmen, welche wir durch raumakustische Simulationen erstellten. Hierfür wurde das Programm *CATT-Acoustic* verwendet. Abb. 5 zeigt den Raum. Der Raum hat ungefähr die Abmessungen $20\,\mathrm{m} \times 20\,\mathrm{m} \times 6\,\mathrm{m}$, wobei einzelne Flächen leicht schräg angeordnet sind, um Raummoden und Echos zu vermeiden. Das Raumvolumen beträgt ca. $2600\,\mathrm{m}^3$.

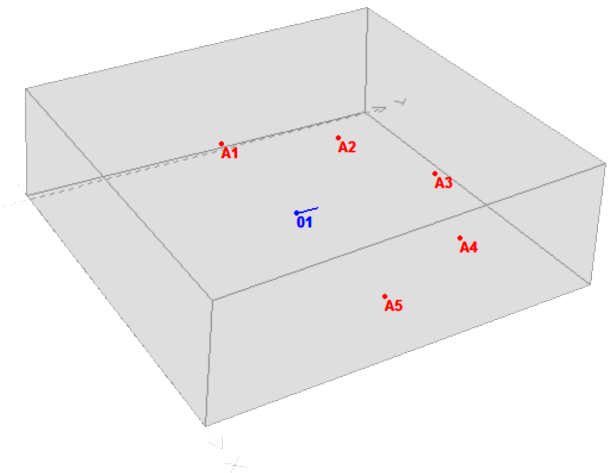


Abbildung 5 – Virtueller Raum mit Empfängerposition und fünf Senderpositionen.

In der Mitte des Raumes ist die Empfängerposition in blau eingezeichnet. Um diese herum sind rot die fünf Senderpositionen eingezeichnet. Die Sender haben allesamt einen Abstand von 8 m zum Empfänger. Sowohl Sender als auch Empfänger befinden sich auf

einer Höhe von 1.5 m über dem Boden. Von diesen fünf Sender-Empfänger Strecken wurde jeweils eine 16-kanalige Ambisonics-Impulsantwort (dritte Ordung) berechnet.

Um Impulsantworten mit weniger Nachhall zu erhalten, wurden die Impulsantworten der Raumsimulation mit einer abklingenden Exponentialfunktion gewichtet, um den Diffusanteil zu verringern. Bei den ursprünglich simulierten Impulsantworten befand sich der Empfänger deutlich außerhalb der Hallradii der Empfänger. Mit der exponentielle Gewichtung wird durch künstliche Verkürzung der Nachhallzeit der Hallradius vergrößert, was zu einem geringerem Diffusanteil in den Impulsantworten führt. Um akustische Szenen zu erstellen standen uns also zum einen Impulsantworten zur Verfügung, bei denen der (virtuelle) Empfänger außerhalb des Hallradius positioniert war und zum anderen Impulsantworten, bei denen er im Hallradius positioniert war.

Die Impulsantworten wurden nun benutzt, um Audiomaterial mit ihnen zu falten, somit Sprecher bzw. Instrumente im Raum zu positionieren und akustische Szenen zu erstellen. Mit den beiden Arten von Impulsantworten (Empfänger innerhalb des Hallradius, Empfänger außerhalb des Hallradius) generierten wir jeweils zwei Szenarien (ein Sprachszenario und ein Musikszenario), was uns zu den o.g. vier simulierten Ausgangsaufnahmen führte.

Die Aufstellung der Sprecher bzw. Instrumente wurde wie folgt gewählt:

Sprachszenario

Zwei weibliche und ein männlicher Sprecher tragen gleichzeitig unterschiedliche Texte vor (s. Abb. 6).

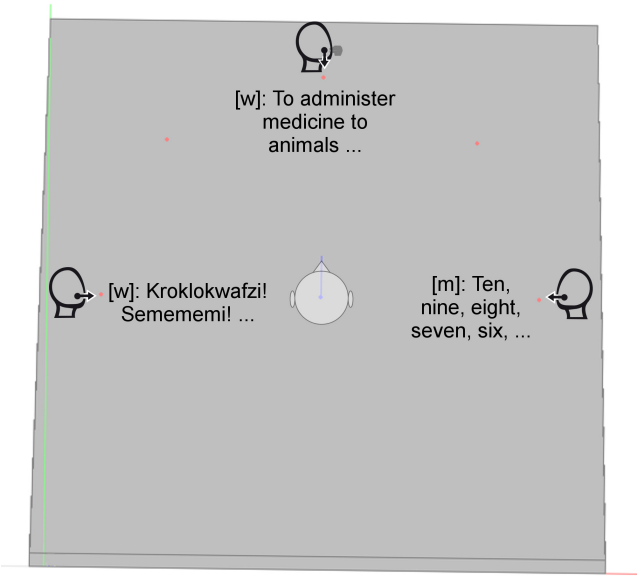


Abbildung 6 – Anordnung der Sprecher.

Musikszenario

Vier ausgewählte Instrumentenspuren (gleichzeitig spielender Instrumente) aus dem Song „Reckoner“ der Band „Radiohead“ wurden im Raum platziert (siehe Abb. 7).

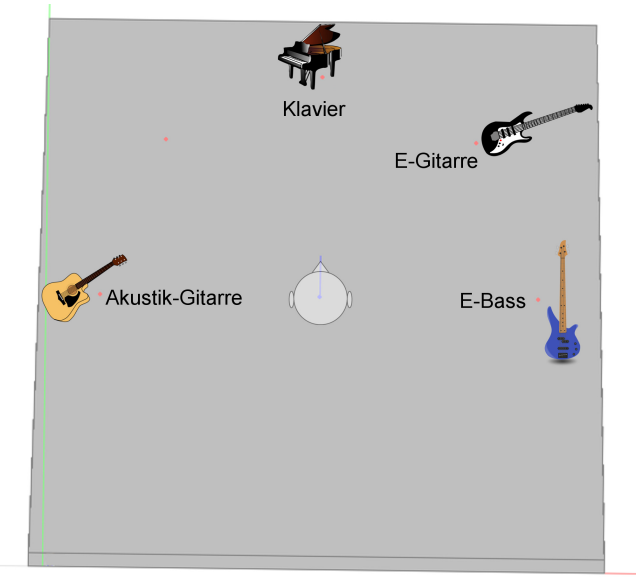


Abbildung 7 – Anordnung der Instrumente.

Die erstellten Ambisonics-Aufnahmen dritter Ordnung wurden als unbeeinträchtigte Referenzsignale im Sinne der ITU-Empfehlung *ITU-R BS.1534-3* [7] verwendet. Wenn man ausschließlich die ersten vier Kanäle (W, X, Y, Z) verwendet, erhält man ein Ambisonics-Signal erster Ordnung (B-Format), so wie es z.B. auch mit dem Soundfield-Mikrofon hätte entstehen können. Diese Aufnahmen erster Ordnung dienten als Ausgangsmaterial für das DirAC-Verfahren, mit welchem nun versucht wurde, die räumliche Auflösung wieder auf Ambisonics dritter Ordnung zu erhöhen.

Das Verfahren wurde jeweils mit fünf unterschiedlichen FFT-Längen (16, 64, 256, 1024 und 4096 Punkte) angewandt. Die Mittelungskonstanten haben wir nach Gehör und unter den gleichen Abhörbedingungen wie im Hörversuch so eingestellt, dass Artefakte möglichst gut vermieden werden. Wir empfanden, dass dies am besten bei einer Mittelung von 20 ms, sowohl für Richtung und Diffusität und unabhängig von der FFT-Länge, funktionierte.

In getrennten Testdurchläufen sollten im Hörexperiment pro simulierter Ausgangsaufnahme das Referenzsignal, die unverarbeitete Aufnahme erster Ordnung und die fünf DirAC-verarbeiteten Signale (also insgesamt sieben Testsignale) miteinander verglichen werden.

Die Generierung der verschiedenen Testsignale ist in Abbildung 8 in Bezug auf den Hörversuch noch einmal graphisch dargestellt.

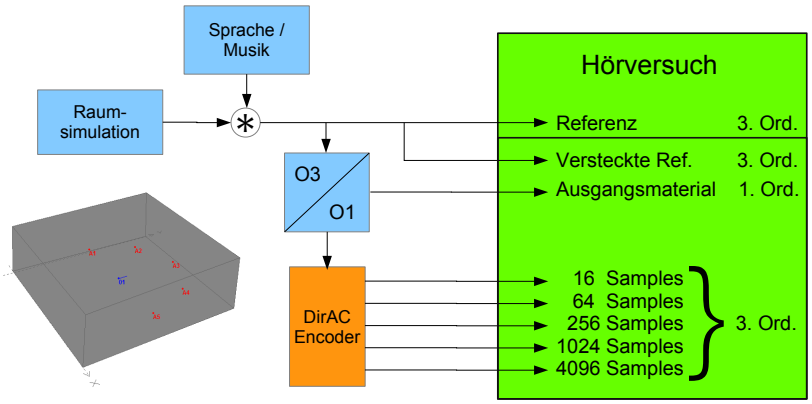


Abbildung 8 – Generierung der Testsignale und Verwendung im Hörversuch.

3.2.2 Interface

Für die Wiedergabe und Bewertung der Audiobeispiele wurde in Python mithilfe der Library *Tkinter* ein Testinterface implementiert, das sich im wesentlichen am MUSHRA-Test [7] orientiert. Da die Audiobeispiele mit REAPER wiedergegeben wurden, wurde zur Kommunikation zwischen dem Interface und REAPER das OSC-Protokoll (Open Sound Control) verwendet. Die von den Testpersonen abgegebenen Bewertungen wurden in einer Textdatei gespeichert.

Das Interface ermöglicht das Abspielen von Audiobeispielen und das Umschalten zwischen unterschiedlichen Kodierungen (DirAC mit 16–4096 Punkte-FFT, Referenz bzw. 3. Ordnung, Ausgangsmaterial bzw. 1. Ordnung). Wie bereits erwähnt, werden sieben Kodierungen miteinander verglichen. Jede dieser Kodierungen wird graphisch durch einen Bewertungsfader und einen *Play/Stop*-Knopf repräsentiert (siehe Abb. 9).

Die Anordnung der Signale auf der grafischen Oberfläche wird in jedem Durchgang neu randomisiert. Die Versuchsperson weiß also nicht, welche getestete Kodierung durch welchen Fader/Knopf repräsentiert wird. Die Beschriftungen der einzelnen Klangbeispiele dienen nur zur Orientierung der Versuchsperson und geben keinen Hinweis auf die wirkliche Signalreihenfolge.

Das Referenzsignal befindet sich einmal versteckt unter den Testsignalen, kann aber von der Versuchsperson auch gezielt abgespielt werden. Die Audiobeispiele werden ab dem Drücken eines der *Play*-Knöpfe in Endlosschleife wiedergegeben. Nun kann durch Drücken eines anderen *Play*-Knopfes ohne Anhalten der Wiedergabe zu einer anderen Kodierung gewechselt werden. Die Möglichkeit, an charakteristischen Stellen zu einer anderen Kodierung zu wechseln, erlaubt einen zügigen Vergleich. Durch Drücken des *Stop*-Knopfes beim aktuell abgespielten Testsignal kann die Wiedergabe angehalten werden.

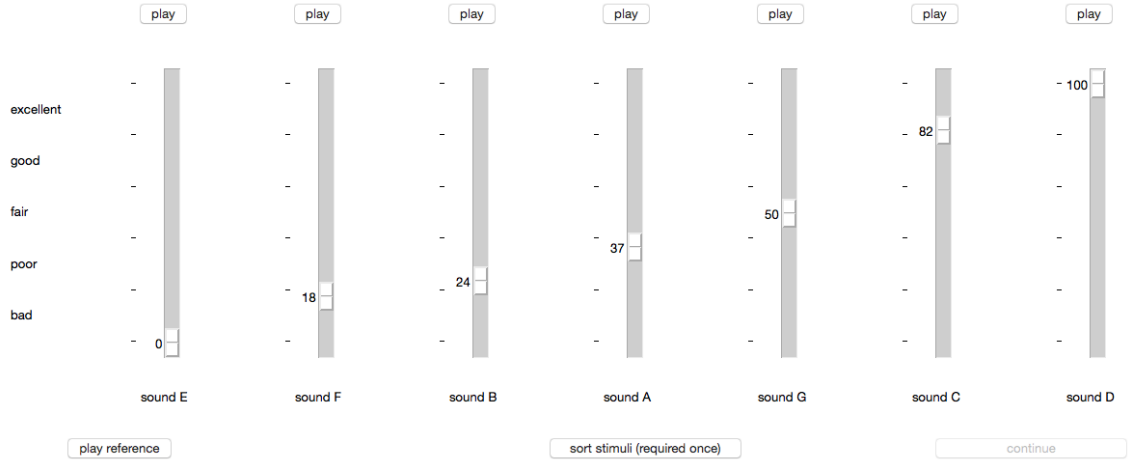


Abbildung 9 – Software-Interface für den Hörversuch.

Eine nützliche Besonderheit der Testoberfläche ist die Sortierfunktion. Die Testperson kann die Testsignale, nachdem sie diese einmal bewertet hat, von rechts nach links in der Reihenfolge der Bewertung sortieren. Die Beschriftungen ändern sich entsprechend der Sortierung. Die Sortierfunktion kann beliebig oft genutzt werden. In diesem Versuch wurde zusätzlich verlangt, dass die Versuchsperson mindestens einmal die Testsignale sortiert und jedes Signal inklusive der Referenz mindestens einmal angehört hat, bevor sie zum nächsten Testdurchlauf wechseln konnte.

Wenn die Versuchsperson den gesamten Test abgeschlossen hat, wurden die Versuchsleiter per Email benachrichtigt.

3.2.3 Aufbau und Durchführung

Der Hörversuch wurde im IEM-CUBE durchgeführt. Die Lautsprecheranordnung war die CUBE-Halbkugel mit 24 Lautsprechern, in deren Zentrum sich die Testperson an einem Tisch sitzend befandt. Das Testinterface wurde an einem Laptop bedient. Die Testsignale im Ambisonics-Format wurden für die Lautsprecheranordnung mit dem am IEM entwickelten *AllRAD*-Ambisonics-Decoder [6] dekodiert.

Abb. 10 zeigt eine Versuchsperson beim Bewerten der Testsignale.

Zu Beginn des Tests wurde den Versuchspersonen die Funktionsweise des Interfaces, die Bewertungskriterien und der Versuchsablauf erklärt, welcher in Abb. 11 veranschaulicht ist. Es folgte eine kurze Lernphase, in der die Versuchsperson das Interface kennenlernen und Fragen an die Versuchsleiter stellen konnte. Hierbei sollte im ersten Trainingsdurchlauf ein Beispiel bezüglich Lokalisationsschärfe, dann im zweiten ein Beispiel bezüglich Audioqualität bewertet werden.

Nach einer kurzen Pause führten wir die erste Session des Tests durch. Hier sollte von der Versuchsperson die Lokalisationsschärfe bei den sieben Kodierungen bewertet werden. Den Teilnehmern wurde, um einen Bias durch den Unterschied erfahrener und unerfahrener Hörer zu vermeiden, mitgeteilt, dass das Referenzsignal einmal unter den zu

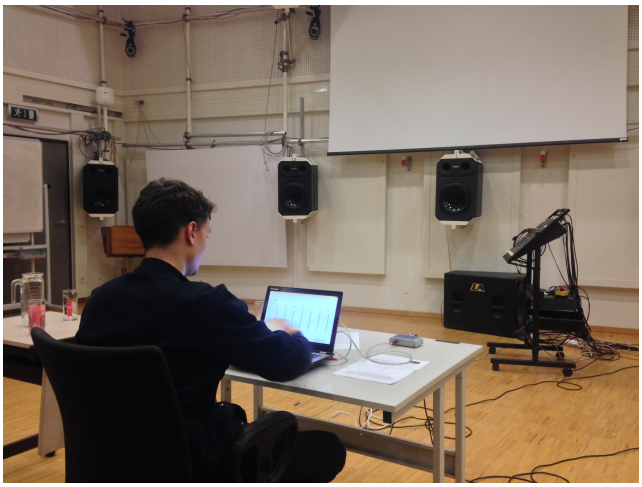


Abbildung 10 – Versuchsperson an der Abhörposition.

bewertenden Kodierungen versteckt war. In vier getrennten Testdurchläufen wurde jeweils eines der akustischen Szenarien präsentiert. Die Reihenfolge der akustischen Szenarien wurde randomisiert.

Analog zur Bewertung der Lokalisationsschärfe sollten die Testteilnehmer nach einer weiteren kurzen Pause in einer zweiten Test-Session die gleichen Beispiele bezüglich Audioqualität bewerten. Die Reihefolge der akustischen Szenarien war erneut randomisert.

Wir ließen die Teilnehmer immer zuerst die Lokalisationsschärfe bewerten, da dies unserer Einschätzung nach die anspruchsvollere Aufgabe darstellte. Dies wurde von den Teilnehmern bestätigt. Die Bewertung der Lokalisationsschärfe mit bereits ermüdeten Ohren vermutlich schwieriger gewesen.

3.3 Ergebnisse

3.3.1 Methoden

Das durchgeführte Experiment kann in zwei Teilexperimente (Bewertung der Audioqualität, Bewertung der Lokalisationsschärfe) mit jeweils einer abhängigen Variable (der Bewertung auf einer Skala von 0 bis 100) und drei Within-Subjects-Faktoren, nämlich

- Audiokodierung (7 Stufen: die verschiedenen FFT-Längen),
- Szene (2 Stufen: Sprache oder Musik)
- und Diffusanteil (2 Stufen: Empfänger innerhalb oder außerhalb des Hallradius),

aufgeteilt werden.

Da an unserem Experiment 7 Personen teilnahmen, wurden insgesamt für die abhängige Variable pro Teilexperiment $7 \cdot 7 \cdot 2 \cdot 2 = 196$ Werte gesammelt. Wir

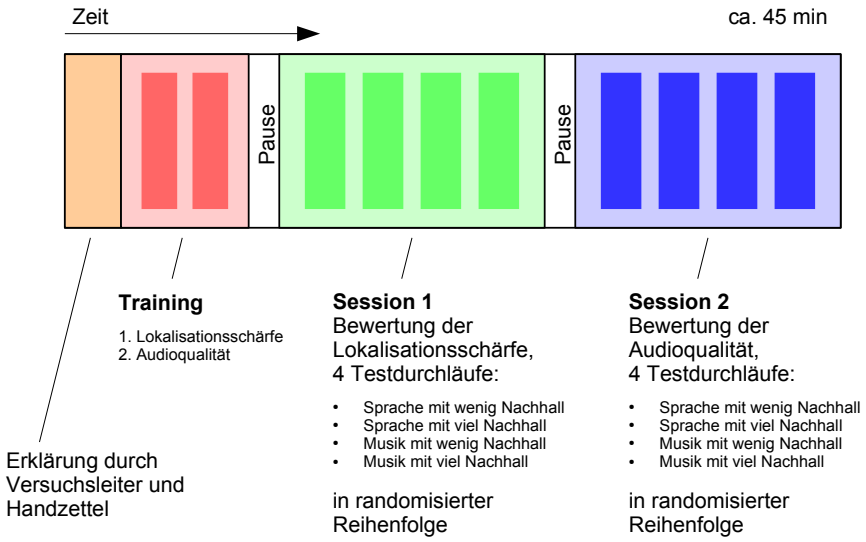


Abbildung 11 – Versuchsablauf: Erklärung, Training und zwei Sessions.

entschieden uns dazu, den Versuch mit Varianzanalysen und nachfolgenden Tukey-HSD-Mehrfachvergleichstests auszuwerten.

Da eine dreifaktorielle Varianzanalyse und vor allem die eventuell festzustellenden Interaktionseffekte schwer auszuwerten ist und ggf. auch keine angemessenen Mehrfachvergleichs-Tests existieren, teilten wir jedes der Teilexperimente in vier einfaktorielle Varianzanalysen auf, bei denen der Faktor die Audiokodierung war. Für diese Analysen wurde die beiden Nullhypothesen H_{01} und H_{02} (H_{01}: „Die Audiokodierung hat keinen Einfluss auf die wahrgenommene Audioqualität“, H_{02}: „Die Audiokodierung hat keinen Einfluss auf die wahrgenommene Lokalisationsschärfe“) überprüft. Da es sich um ein Pilotexperiment handelt, wurde das Signifikanzniveau bei den Mehrfachvergleich-Tests auf $\alpha = 10\,\%$ festgelegt.

3.3.2 Analyse

Im Folgenden werden die Auswertungen der Bewertung der Audioqualität und der Lokalisationsschärfe jeweils zusammengefasst.

Für jedes Szenario sind die ANOVA-Tabellen und die graphische Darstellung der Mittelwerte und der *Vergleichsintervalle*[1] [8] angegeben. In den Abbildungen 12 bis 15 sind jeweils eine Kurve für Audioqualität und Lokalisationsschärfe gegenübergestellt. Die

1. Hierbei handelt es sich nicht um Konfidenzintervalle, sondern um Intervalle, die unter Berücksichtigung des ANOVA-Modells und der Abhängigkeit der Stichproben gleicher Versuchspersonen festlegen, ob ein signifikanter Unterschied gegeben ist oder nicht. Überschneiden sich zwei Intervalle, so sind die Mittelwerte der jeweiligen Gruppen nicht signifikant unterschiedlich; überschneiden sie sich nicht, so sind sie es.

Bewertungen der DirAC-Kodierungen sind als verbundene Datenpunkte und Balken für die Vergleichsintervalle eingezeichnet; die Bewertungen des Ausgangsmaterials sind als horizontale Linien und halbtransparente Flächen für die Vergleichsintervalle dargestellt. Die Bewertungen der versteckten Referenzsignale gingen nicht in die Analyse mit ein und sind in diesen Abbildungen nicht dargestellt, da diese in der Regel mit 100 Punkten bewertet wurde. Allerdings konnten in einigen Fällen festgestellt werden, dass Versuchspersonen andere Testsignale und das versteckte Referenzsignal verwechselten.

Sprache mit geringem Diffusanteil

Wie den Tabellen 1 und 2 entnommen werden kann, können beide Nullhypothesen abgelehnt werden.

Quelle	SS	df	MS	F	p(F)
Kodierung	34600.0	6	5766.7	21.93	0.000
Testperson	1629.1	6	271.5	1.03	0.420
Residuum	9465.2	36	262.9	-	-
Total	45694.3	48	-	-	-

Tabelle 1 – Varianzanalyse der Bewertung der Audioqualität (Sprachszenario, geringer Diffusanteil).

Quelle	SS	df	MS	F	p(F)
Kodierung	31921.6	6	5320.3	13.24	0.000
Testperson	4227.9	6	704.7	1.75	0.137
Residuum	14466.9	36	401.9	-	-
Total	50616.5	48	-	-	-

Tabelle 2 – Varianzanalyse der Bewertung der Lokalisationsschärfe (Sprachszenario, geringer Diffusanteil).

Tukey-Mehrfachvergleichstests wurden durchgeführt. Abb. 12 zeigt die Randmittelwerte Vergleichsintervalle ($\alpha = 10\,\%$). Wie zu sehen ist, wurde die Lokalisationsschärfe für alle FFT-Längen oberhalb von 16 Samples signifikant verbessert. Im mittleren Bereich wurde auch die Audioqualität tendenziell verbessert, wohingegen z.B. bei 16 Samples FFT-Länge eine signifikante Verschlechterung festgestellt wurde.

Als die Versuchspersonen auf die Audioqualität achten sollten, kam es einmal zu einer Verwechsung zwischen dem versteckten Referenzsignal und DirAC-kodiertem Signal mit einer FFT-Länge von 1024 Samples und einmal zwischen Referenzsignal und DirAC-kodiertem Signal mit einer FFT-Länge von 4096 Samples. Lag der Fokus auf der Lokalisierbarkeit, so wurde das DirAC-Signal mit einer FFT-Länge von 1024 Samples einmal mit der versteckten Referenz verwechselt.

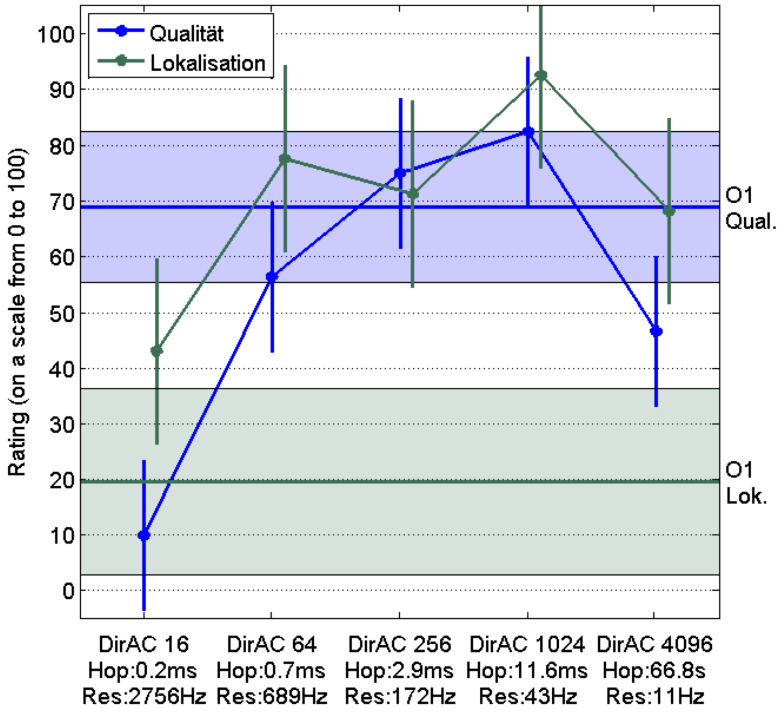


Abbildung 12 – Randmittelwerte und Vergleichsintervalle (Sprachszenario, geringer Diffusanteil).

Sprache mit großem Diffusanteil

Die Tabellen 3 und 4 zeigen die Ergebnisse der Varianzanalyse. Auch hier können beide Nullhypothesen abgelehnt werden.

Quelle	SS	df	MS	F	p(F)
Kodierung	33300.0	6	5550.0	23.65	0.000
Testperson	3339.1	6	556.5	2.37	0.049
Residuum	8446.6	36	234.6	-	-
Total	45085.7	48	-	-	-

Tabelle 3 – Varianzanalyse der Bewertung der Audioqualität (Sprachszenario, großer Diffusanteil).

Quelle	SS	df	MS	F	p(F)
Kodierung	30655.1	6	5109.2	17.81	0.000
Testperson	4736.6	6	789.4	2.75	0.026
Residuum	10326.3	36	286.8	-	-
Total	45718.0	48	-	-	-

Tabelle 4 – Varianzanalyse der Bewertung der Lokalisationsschärfe (Sprachszenario, großer Diffusanteil).

Abb. 13 zeigt die Randmittelwerte und Vergleichsintervalle. Wiederum wurde die Lokalisationsschärfe signifikant verbessert. Die Qualität wurde bei 16, 64 und 4096 Samples FFT-Länge signifikant verschlechtert. Eine tendenzielle Verbesserung kann bei keiner FFT-Länge festgestellt werden.

Als die Versuchspersonen sich auf die Lokalisierbarkeit konzentrierten, kam es einmal zu einer Verwechslung zwischen Referenzsignal und DirAC mit 1024-Punkte-FFT.

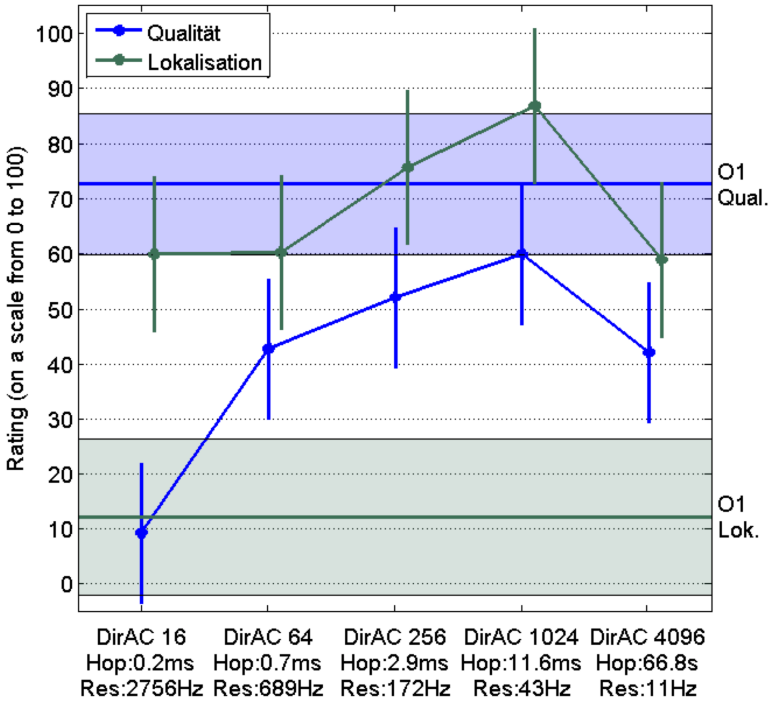


Abbildung 13 – Randmittelwerte und Vergleichsintervalle (Sprachszenario, großer Diffusanteil).

Musik mit geringem Diffusanteil

Wie erwartet können die Nullhypothesen auch hier abgelehnt werden (s. Tabelle 5 und 6).

Quelle	SS	df	MS	F	p(F)
Kodierung	52987.0	6	8831.2	29.24	0.000
Testperson	1165.3	6	194.2	0.64	0.695
Residuum	10874.2	36	302.1	-	-
Total	65026.4	48	-	-	-

Tabelle 5 – Varianzanalyse der Bewertung der Audioqualität (Musikszenario, geringer Diffusanteil).

Quelle	SS	df	MS	F	p(F)
Kodierung	48267.3	6	8044.6	29.17	0.000
Testperson	1812.2	6	302.0	1.10	0.384
Residuum	9928.4	36	275.8	-	-
Total	60007.9	48	-	-	-

Tabelle 6 – Varianzanalyse der Bewertung der Lokalisationsschärfe (Musikszenario, geringer Diffusanteil).

Abb. 14 zeigt die Randmittelwerte und Vergleichsintervalle. Der Darstellung kann entnommen werden, dass ab einer FFT-Länge von 256 Samples die Lokalisationsschärfe signifikant verbessert wurde. Bei kurzen FFT-Längen zeigen sich starke Qualitätseinbußen durch Artefakte. Die Audioqualität konnte nicht verbessert werden.

Bei Fokussierung auf die Audioqualität wurde einmal das DirAC-Signal mit 1024-Punkte-FFT mit dem versteckten Referenzsignal verwechselt. Bei Fokussierung auf die Lokalisierbarkeit wurde einmal das DirAC-Signal mit 256-Punkte-FFT und einmal das DirAC-Signal mit 4096-Punkte-FFT mit dem Referenzsignal verwechselt.

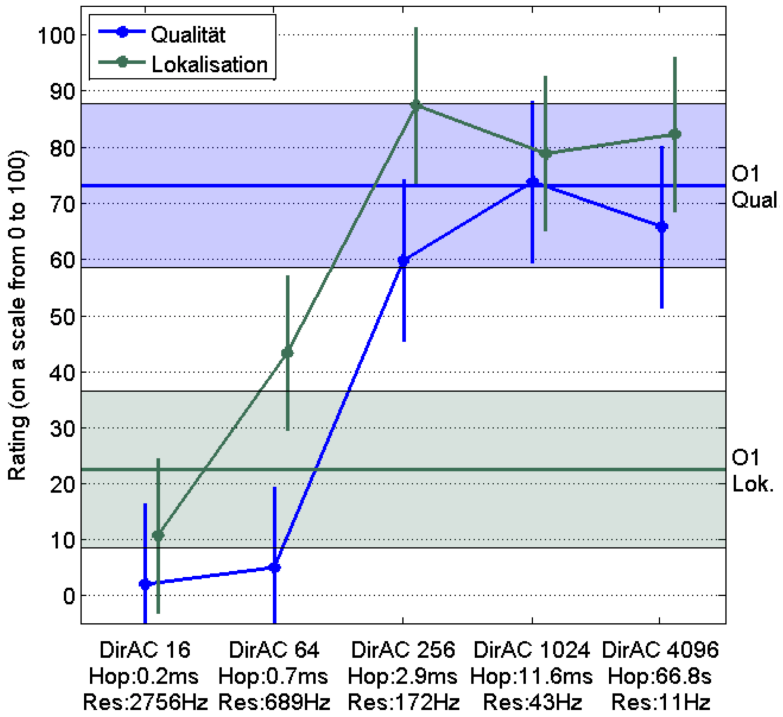


Abbildung 14 – Randmittelwerte und Vergleichsintervalle (Musikszenario, geringer Diffusanteil).

Musik mit großem Diffusanteil

Auch hier können beide Nullhypothesen abgelehnt werden (s. Tabellen 7 und 8).

Quelle	SS	df	MS	F	p(F)
Kodierung	54796.8	6	9132.8	41.28	0.000
Testperson	2301.7	6	383.6	1.73	0.141
Residuum	7964.3	36	221.2	-	-
Total	65062.8	48	-	-	-

Tabelle 7 – Varianzanalyse der Bewertung der Audioqualität (Musikszenario, großer Diffusanteil).

Quelle	SS	df	MS	F	p(F)
Kodierung	36116.4	6	6019.4	12.98	0.000
Testperson	4371.3	6	728.5	1.57	0.184
Residuum	16697.3	36	463.8	-	-
Total	57185.0	48	-	-	-

Tabelle 8 – Varianzanalyse der Bewertung der Lokalisationsschärfe (Musikszenario, großer Diffusanteil).

Abb. 15 zeigt die Randmittelwerte und Vergleichsintervalle. In diesem Fall wurde das schlechteste Ergebnis erzielt. Eine tendezielle Verbesserung der Lokalisationsschärfe konnte zwischen 256 und 4096 Samples erzielt werden. Die Audioqualität wurde in jedem Fall signifikant verschlechtert.

Bei Fokussierung auf die Audioqualität wurde einmal das Referenzsignal mit dem Ausgangssignal (1. Ordnung) verwechselt.

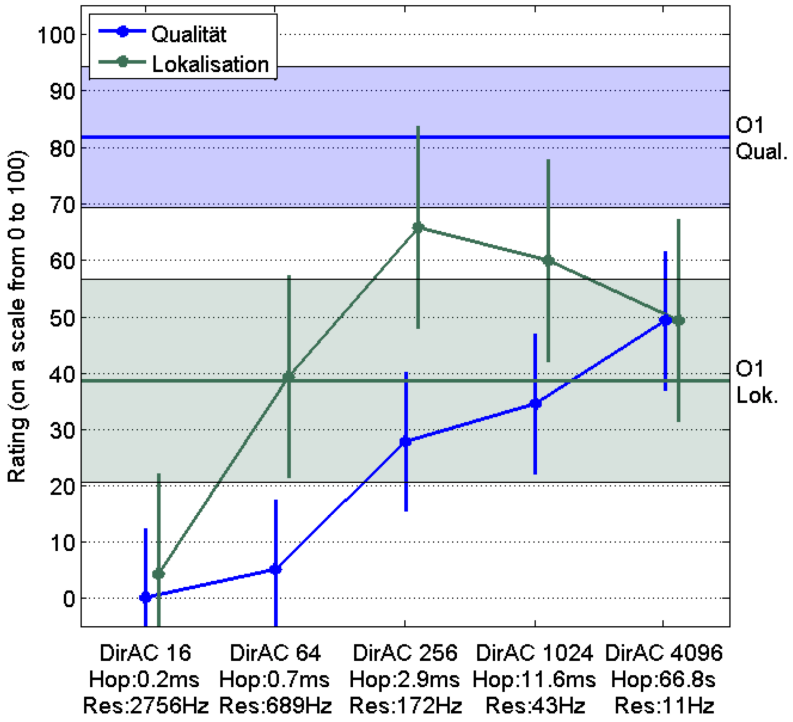


Abbildung 15 – Randmittelwerte und Vergleichsintervalle (Musikszenario, großer Diffusanteil).

3.4 Diskussion

Das Experiment hat gezeigt, dass die Performance des Verfahrens stark signalabhängig ist. Ein dichteres Frequenzspektrum und stärkerer Nachhall beeinträchtigen sowohl Lokalisation als auch Qualität.

Die Bewertung des Verfahrens muss anwendungsabhängig erfolgen: Bei der Sprachübertragung kann, eine Verbesserung der Lokalisation vorausgesetzt, eine geringe Beeinträchtigung der Klangqualität durchaus in Kauf genommen werden. Mit diesem Gedanken kann gefolgert werden, dass die Anwendung des Verfahrens bei Sprache auch bei vergleichsweise halligen Signalen Sinn macht. Am besten schnitt der Algorithmus bei Sprache mit kurzem Nachhall und einer FFT-Länge von 1024 Punkten ab.

Im Falle von Musik sind Klangverschlechterungen in der Regel nicht akzeptabel. Hier konnte nur bei Musik mit wenig Nachhall bei einer FFT-Länge von 1024 Punkten gezeigt werden, dass bei einer Verbesserung der Lokalisierung keine tendenzielle Verschlechterung der Audioqualität auftrat. Der von uns verwendete Algorithmus ist also für realistische Musikaufnahmen nicht geeignet. Gegebenenfalls kann der Algorithmus durch Feineinstellung und weitere Ansätze, die in Kap. 4 vorgestellt werden, verbessert werden.

4 Zusammenfassung und Ausblick

In dieser Arbeit wurde der DirAC-Algorithmus, ein Verfahren zur Auflösungsverbesserung von Surround-Aufnahmen, untersucht. Nach der Behandlung der theoretischen Grundlagen (Kap. 2.1 und 2.2) wurde unsere Implementierung des Verfahrens vorgestellt (Kap. 2.3). Bei der Realisierung gibt es einen gewissen Spielraum. Dies betrifft nicht nur die Herangehensweise bei der Bestimmung von Diffusität und Richtung, sondern auch die Wahl des Wiedergabeverfahrens, das diese Parameter verwendet, um ausgehend vom ursprünglichen Tonmaterial das aufgenommene Schallfeld zu resynthetisieren.

Unsere Implementierung beruht auf der Verwendung der *Short-time Fourier Transformation* und der Berechnung von Diffusität und Richtung im Frequenzbereich. Die Wiedergabe erfolgte mit Ambisonics. Ausgehend von B-Format-Aufnahmen (Ambisonics 1. Ordnung) wurden auf diese Weise Ambisonics-Signale höherer Ordnung erzeugt.

Der Effekt der FFT-Länge wurde in einem Hörversuch untersucht (Kap. 3). Dieser wurde im IEM CUBE mit sieben Teilnehmen durchgeführt. Es zeigten sich große Unterschiede bei den erzielten Ergebnissen. Die Erhöhung des Diffusanteils im Eingangssignal führte zu stärkeren Artefakten und schlechterer Lokalisierbarkeit. Die besten Ergebnisse wurden mit vergleichsweise trockenen Sprachsignalen erzielt, während bei einem halligen Musiksignal kaum eine Verbesserung der Lokalisierbarkeit und außerdem eine Verschlechterung der Klangqualität im Vergleich zum B-Format-Signal beobachtet wurde. Die optimale FFT-Länge lag in den meisten Fällen zwischen 256 und 1024 Punkten.

Da, wie bereits erwähnt, verschiedene Varianten der Implementierung möglich sind und wir zudem nur Gelegenheit zu einer Studie im kleinen Rahmen hatten, erheben wir zwar keinen Anspruch darauf, dass unsere Ergebnisse uneingeschränkt für das DirAC-Verfahren gültig sind. Jedoch liefern die Ergebnisse für die von uns gewählte Implementierung qualitative Aussagen über die Funktionalität des Algorithmus bei Variation des Tonmaterials und des Nachhalls.

Es wurde gezeigt, dass der Algorithmus, vor allem bei einem großen Diffusanteil im Eingangssignal, teilweise nicht optimal funktioniert. Es kommen verschiedene Ansätze in Betracht, um den Algorithmus weiter zu optimieren, die in dieser Arbeit nicht untersucht wurden.

Eine Möglichkeit wäre das Hinzumischen des Originalsignals zum Ausgangssignal. Das Ziel dieser Methode ist die Maskierung der erzeugten Artefakte bei gleichzeitiger Erhaltung des Gewinns an Lokalisationsschärfe durch die DirAC-Verarbeitung.

Die tatsächliche Zerlegung des Eingangssignal in den enthaltenen Direkt- und Diffusanteil, wie in [9] beschrieben, ist ein weiterer möglicher Ansatz. Mit dieser Methode könnte möglicherweise mit dem aus dem Eingangssignal berechneten Direktanteil eine präzisere Richtungsbestimmung erfolgen und gleichzeitig ein authentisches Diffussignal generiert werden.

Der Effekt von ausgewählten Überblendkurven zwischen Direkt- und Diffusanteil könnte untersucht werden.

Darüber hinaus wäre es interessant, die Auswirkungen einer gehörangepassten Filterbank

oder Kurzzeittransformation anstelle der STFT zu untersuchen. Bei der STFT sind alle Frequenzbänder gleich breit, während das menschliche Gehör bei niedrigen Frequenzen eine höhere Auflösung besitzt als bei hohen.

Eine Echtzeitimplementierung des Verfahrens wäre interessant, da es nur so für Internetkonferenzanwendungen geeignet ist.

Für die Evaluierung der Ansätze zur Verbesserung des Verfahrens wäre es wünschenswert, einen Hörversuch mit mehr Teilnehmern durchzuführen und so die Aussagekraft der Ergebnisse zu erhöhen. Es wäre außerdem interessant, Signalcharakteristika wie Diffusität oder spektrale Dichte mithilfe von objektiven Kriterien quantitativ zu erfassen und gemeinsam mit dem Faktor Fensterlänge in Zusammenhang mit subjektiven Empfindungen zu bringen.

Es gibt also durchaus noch Potential für Verbesserungen des Algorithmus, die in folgenden Arbeiten behandelt werden können.

Literatur

[1] V. Pulkki, "Spatial sound reproduction with directional audio coding," *J. Audio Eng. Soc.*, vol. 55, no. 6, pp. 503–516, 2007.

[2] V. Pulkki, M. Laitinen, J. Vilkamo, J. Ahonen, T. Lokki, and T. Pihlajamäki, "Directional audio coding – perception-based reproduction of spatial sound," in *Int. Workshop on the Principles and Applications of Spatial Hearing*, Miyagi, Japan, 2009.

[3] J. Blauert and J. Braasch, "Räumliches Hören," in *Handbuch der Audiotechnik*, S. Weinzierl, Ed. Berlin, Heidelberg: Springer-Verlag, 2008, ch. 3, pp. 87–121.

[4] J. Ahonen and V. Pulkki, "Diffuseness estimation using temporal variation of intensity vectors," in *Applications of Signal Processing to Audio and Acoustics WASPAA'09*. IEEE, 2009, pp. 285–288.

[5] M. Kronlachner, "Plug-in suite for mastering the production and playback in surround sound ambisonics," *Gold-Awarded Contribution to AES Student Design Competition*, 2014.

[6] F. Zotter and M. Frank, "All-round ambisonic panning and decoding," *J. Audio Eng. Soc*, vol. 60, no. 10, pp. 807–820, 2012.

[7] *Method for the subjective assessment of intermediate quality level of audio systems*, ITU-R BS.1534-3, 2015.

[8] Y. Hochberg and A. Tamhane, *Multiple Comparison Procedures. New York:* Wiley.

[9] N. Epain and C. Jin, "Super-resolution sound field analysis," *Cutting Edge in Spatial Audio*, p. 26, 2013.