Seminararbeit aus Algorithmen in Akustik und Computermusik 2

Untersuchung verschiedener Upmixingmethoden für DirAC

Manuel Planton, BSc
Michael Hirschmugl, BSc

Betreuung: Dr. Franz Zotter, Dr. Matthias Frank

Graz, WS 2019/2020

institut für elektronische musik und akusti

Zusammenfassung

Diese Seminararbeit behandelt verschiedene Ausführungen und auch Optimierungen des DirAC-Verfahrens (Directional Audio Coding). Mit diesem Verfahren können Ambisonics Aufnahmen erster Ordnung in ihrer räumlichen Darstellung verbessert und auf eine höhrere Ordnung hochgerechnet werden. Da die Aufnahme in Ambisonics erster Ordnung mit Hilfe von Soundfield Mikrofonen sehr einfach und günstig durchgeführt werden kann, jedoch eine wenig überzeugende Darstellung von Räumlichkeit und Klang bietet, kann ein solches Upmixing die Wiedergabe in vielerlei Hinsicht verbessern. Zu diesem Zweck sollen verschiedene Kombinationen von Upmixing-Methoden untereinander, und auch mit kommerziellen Plugins verglichen werden. Somit wird in dieser Seminararbeit die optimale Realisierung mit dem größten klanglichen Vorteil gesucht.

Inhaltsverzeichnis

1 Einleitung

einleitungstext

2 DirAC

2.1 Annahmen aus der Psychoakustik

text test [FZ19]

2.2 Funktionsweise

text

2.3 Implementierung

In diesem Kapitel wird beschrieben, wie der DirAC-Algorithmus in diesem Projekt implementiert wurde. Die Implementierung liegt als Octave-Skript vor und nutzt das *signal*-Package.

Zu Beginn des Skriptes wird das Eingangssignal im B-Format erster Ordnung eingelesene. Dies muss als *.wav*-Datei vorliegen und genau vier Kanäle umfassen. Die Samplerate wird aus der Datei ermittelt und als Variable *fs* im Skript gespeichert. Anschließend wird eine VBAP-Matrix, je nach vorgegebener Lautsprecherandordnung erzeugt. Diese Matrix wird in späterer Folge verwendet, um das Eingangssignal im B-Format, auf eine bestimmte Lautsprecheranordnung zu dekodieren. Es handelt sich dabei also um eine nicht-parametrische Dekodierung, die mit einer Anordnung von virtuellen Mikrofonen zu vergleichen ist. Die Lautsprecherpositionen können dabei als sphärische Koordinaten in einer Matrix mit genau Spalten vorliegen, oder auch aus einer Textdatei mit kartisischen Koordinaten eingelesen werden. Diese Textdatei muss eine Matrix mit genau drei Spalten (Raumdimensionen) enthalten, wobei diese Spalten untereinander angeordnet sein müssen. Es befinden sich demnach genau die dreifache Anzahl der Lautsprecher als Zeilen in dieser Datei. Die Koordinaten sind dabei Meterangaben als Dezimalwerte. Das Einlesen von kartisischen Koordinaten ist speziell für die Dekodierung auf größere T-Designs äußerst nützlich.

Die vier Kanäle $w[n]$, $x[n]$, $y[n]$ und $z[n]$ des B-Format Eingangssginals werden anschließend in einer Schleife in zeitlichen Blöcken (Fenster) verarbeitet. Ein Fenster besteht dabei aus jeweils 512 Samples und wird in einer 1024-Punkt Fouriertransformation mit der Hilfe der Funktion *fft()* in den Frequenzbereich transformiert. Die FFT-Länge wurde als doppelte Blocklänge gewählt um Aliasing zu vermeiden. Die zeitlichen Fenster werden noch mit einer Hanning-Fensterfunktion für die Resynthese beaufschlagt. Ein Hanning-Fenster kann in Octave mit der Funktion *hanning()* erzeugt werden und einfach als Vektor

mit den Zeitsignalen multipliziert werden.

Analyse von Richtungs- und Diffusanteil Zur Bestimmung von Richtungs- und Diffusanteil werden zunächst Schallschnelle und Energie berechnet. Die Schnallschnelle wird als Vektor $\mathbf{V}_m[k] = [X_m[k], Y_m[k], Z_m[k]]$ aus den gerichteten Anteilen (der Druckgradienten-Mikrofone) des B-Format Signals bestimmt. Der Schalldruck ist schlicht der omnidirektionale Anteil $W_m[k]$. Die Indizes m und k werden hier zur Kennzeichnung des Zeitfensters n als Funktion der Frequenzzahl k verwendet.

Die Schallintensität $I_m[k]$ wird aus dem Schnellevektor und dem konjugiert komplexen Schalldruck in Gleichung 1 abgeleitet, wobei hier rein der Realteil heranzogen wird, da sonst auch die Blindeinteile des Schnellevektors miteinbezogen würden. Der Intensitätsvektor stellt bereits die Schalleinfallsrichtung für alle Frequenzbins einzeln dar, jedoch entgegengesetzt der Einfallsrichtung $\mathbf{D}_m[k]$:

$$-\mathbf{D}_m[k] = \mathbf{I}_m[k] = \Re(W_m[k]^* \cdot \mathbf{V}_m[k]) \tag{1}$$

Die Schallenergie $E_m[k]$ wird mithilfe von Gleichung 2 ausgewertet:

$$E_m[k] = \frac{|W_m[k]|^2 + ||\mathbf{V}_m[k]||^2}{2} \tag{2}$$

Um Sprünge in der Lautsprecherzuordnung von gerichteten Signalen bei raschen Bewegungen zu vermeiden, wird der Intensitätsvektor zusätzlich durch einen Mittelwertbildung geglättet. Diese Glättung kann im Skript mit frequenzabhängiger Zeitkonstante vorgegeben werde. Zusätzlich wird auch der Energievektor geglättet. Anschließend wird der Intensitätsvektor verwendet um die Einfallsrichtung in sphärischen Koordinaten zu bestimmen.

Die Ḋiffusheit$\psi_m[k]$ kann schlussendlich aus dem Vergleich des Betrags des Intensitätsvektors mit der Energie ermittelt werden (Glg. 3). Der Erwartungswert entspricht hier den (zeitlich) gemittelten Vektoren.

$$\psi_m[k] = \sqrt{1 - \frac{||\mathbb{E}(\mathbf{I}_m[k])||}{\mathbb{E}(E_m[k])}} \tag{3}$$

Upmixing auf Lautsprecherandordnung Zur Dekodierung auf eine bestimmte Lautsprecheranordnung wird eine Matrix aus virtuellen Mikrofonen berechnet, woraus sich eine Tabelle mit Gain-Werte ergibt. Diese Tabelle kann die einzelne Frequenzbins in einem VBAP-Verfahren den enstprechenden Lautsprechern zuordnen. In dieser Seminararbeit wurden dabei zwei Ansätze für die Dekodierung verglichen: Einerseits die direkte Dekodierung auf die Lautsprecheranordnung des widergebenden Systems, und andererseits ein eher allgemeiner Ansatz womit die Dekodierung auf ein ambisonisches B-Format Array vierter Ordnung erfolgt. Dieser Dekodierungsansatz hat den Vorteil, dass die Wiedergabeanordnung zum Zeitpunkt des Upmixings nicht vorgegeben sein muss. Auch für die

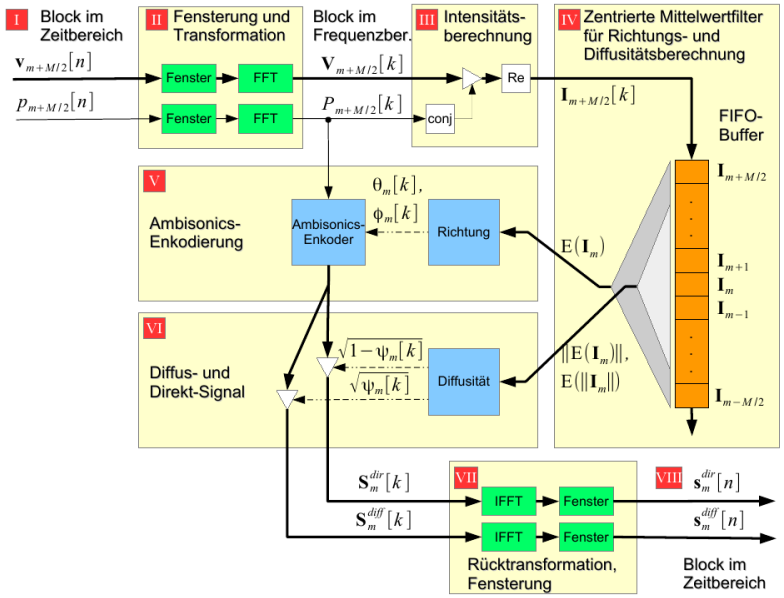

Abbildung 1 – Flussdiagram des DirAC Algorithmus in Octave

Klangqualität stellen sich hierdurch Vorteile heraus welche im Hörversuch besprochen werden.

Die Dekodierung auf eine gegebene Lautsprecheranordnung durch eine VBAP-Tabelle wird hier nicht weiter ausgeführt und kann zum Beispiel im Buch Ämbisonics"**??** nachgelesen werden. Für das B-Format vierter Ordnung wird ein sogenanntes T-Design als virtuelle Anordnung von Lautsprechern verwendet. Ein T-Design ist grundsätzlich zur Darstellung der Maxima von sphärischen Harmonischen notwendig. Wird ein T-Design neunter Ordnung eingesetzt ("9-Design", Abb. 2.3) ergibt sich daraus eine sphärische Anordnung von 48 Lautsprechern, welche die verlustlose Konvertierung zwischen B-Format und einer solchen Lautsprecheranordnung bietet. Es lässt sich sagen, dass ein 9-Design dem ambisonischen B-Format vierter Ordnung entspricht.

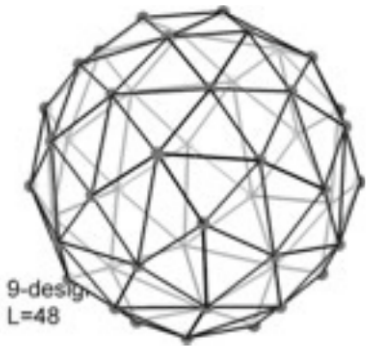

Abbildung 2 – 9-Design T-Design

Zur Dekodierung wird in Octave also eine Matrix mit Lautsprecheramplituden für Schalleinfallsrichtungen erstellt. Somit kann einer gegebenen Richtung (bestehend aus Azi-

muth und Elevation) eine Amplitude am entsprechenden Lautsprecherkanal zugeordnet werden.

Trennung von Diffus- und Richtungsanteil Die Trennung von Diffus- und Richtungsanteil erfolgt im Frequenzbereich, wobei beide Anteile aus dem omindirektionalen Anteil des B-Format Signals durch gewichtete Filterung erzeugt werden. Prinzipiell wird für den gerichteten Anteil ein spektraler Block im Frequenzbereich mit zwei Matrizen multipliziert: Einerseits mit der VBAP-Tabelle um die Richtung einem Lautsprecher zuzuordnen, und andereseits ebenfalls mit dem frequenzabhängigen Diffusitätsvektor. Somit werden Frequenzbins mit frequenzabhängigen Diffusitätswerten gewichtetet und diesen entsprechende Lautsprecheramplituden zugeordnet.

Nach der Berechnung des gerichteten Signalanteils kann auch der Diffusanteil aus dem omnidirektionalen Signal erzeugt werden. Dieser entspricht lediglich dem nicht-gerichteten Anteil des omnidirektionalen Signals und wird daher aus der Filterkurve des Direktanteils bestimmt. Somit ergeben sich an dieser Stelle zwei frequenzabhängige Signalmatrizen für Diffus- und Richtungsanteil. Beide Matrizen besitzen eine Spalte für jeden Lautsprecherkanal und die Zeilenzahl entspricht der FFT-Länge.

Resynthese Die Resynthese entspricht einem gewöhnlichem Overlap-Add Verfahren und erzeugt mittel inverser Fast Fourier Transformation wieder Zeitsignale aus den Matrizen für Richtungs- und Diffusanteil. Im Zeitbereich kann an dieser Stelle weiters auch (optional) die Dekorrelation des Diffusanteils durchgeführt werden. In unserem Skript wurde zur Dekorrelation in Octave ein Random Phase Algorithmus verwende, welcher im Kapitel **??** näher erläutert wird.

2.4 Dekorrelationsverfahren

Feedback Network Delay text

Widening Plugin text

Random Phase text

3 Hörversuch

Die unterschiedlichen Dekodierungsmethoden und Dekorrelationsalgorithmen wurden anschließend in einem Hörversuch auf die Probe gestellt. Hierzu wurden verschiedene Ambisonics Aufnahmen im B-Format herangezogen und mit unterschiedlichen Kombinationen aus Diffusität und Dekodierung bearbeitet. Die Ergebnisse konnten dann direkt im

Produktionsstudio des Instituts für elektronische Musik wiedergegeben und von Versuchspersonen bewertet werden. Der Augenmerk sollte dabei auf Klangqualität und Darstellung der Räumlichkeit liegen. Die folgenden Kapitel sollen die Annahmen, Hypothesen, den Aufbau und die Ergebnisse dieses Versuchs darstellen.

Generell kann der Hörversuch in die folgende Übersicht zerlegt werden:

- Reaper Projekt gesteuert durch Mushra-Test
- Vergleich von 6 Algorithmen + Referenz
- Ausgehend von 4 verglichenen Testsignale
- Einschätzen von Klangqualität und Räumlichkeit

3.1 Ziele des Hörversuchs

Prinzipiell sollten mit dem Hörversuch folgende Forschungsfragen beantwortet werden:

- Bietet das Upmixing durch DirAC auf ein Amisonics-Format höhrer Ordnung einen Vorteil bezüglich der darzustellenden Räumlichkeit der Aufnahme?
- Ist es möglich Upmixing mit DirAC ohne negative Beeinflussung der Klangqualität durchzuführen?
- Welche Kombination aus Dekorrelation und Dekodierung bietet dabei die (subjektiv und objektiv) beste Wiedergabe der Klangqualität?
- Bietet das Upmixing auf ein allgemeines B-Format vierter Ordnung Vorteile bezüglich Klangqualität und Räumlichkeit?
- Wie wird ein kommerzielles Plugin wie "Harpexïm Vergleich zu einer Lösung mit DirAC bewertet?

Weiters soll die Beschaffenheit des Ausgangssignals dabei einbezogen werden. Speziell geht es dabei um die Frage ob eine sehr diffuse Aufnahme ähnlich bewertet wird wie ein eher gerichtetes Signal im B-Format erster Ordnung.

In ersten Versuchen konnten wir die Algorithmen bereits selbst vergleichen und vermuten daher, dass die Lokalisation und Darstellung der Räumlichkeit durch DirAC jedenfalls positiv beeinflusst werden kann, wennauch dies durchaus von der eingesetzten Dekorrelationsmethode stark abhängig ist. Weiters wird angenommen, dass eine bestimmte Kombination von Algorithmen durchaus vergleichbare Ergebnisse mit dem kommerziellen karpex-Plugin bieten könnte. Speziell die Dekodierung auf das T-Design (9-Design) mit 48 virtuellen Schallquellen könnte Vorteile in der Wiedergaben bieten, da die geringen Abstände bei einer derartig dichten Anordnung virtueller Quellen zu einer feinen Granularität führen.

3.2 Aufbau und Konzept

Ausgangssignale Es wurden insgesamt vier unterschiedliche Aufnahmen als Testsignale herangezogen. Dabei wurden auch zwei Aufnahmen im B-Format synthetisch erzeugt, um auch sehr gerichtete Signale vergleichen zu können:

- Synthetisches Zirpen, rotierend in Azimuth

— Synthetisches Zirpen, rotierend in Azimuth, nachträglich verhallt
— Live-Musik Aufnahme
— Umgebungsgeräusche Straßenkreuzung

Die synthetischen Signale wurden dabei mit Hilfe eines weiteren Skriptes in Octave erzeugt. Das zusätzlich verhallte Signal wurde nachträglich in Reaper mit dem FDN Reverb Plugin aus der IEM Plugin Suite bearbeitet. Dies soll einfach einen direkteren Vergleich der Performance zwischen stark gerichtetem und stark diffusem Signal bieten. Die Live-Musik Aufnahme wurde im Cube angefertigt und bietet einen große Räumlichkeit, wobei die Ambience-Aufnahme an einer Straßenkreuzung hier die größte Diffusität aufweist. Diese Aufnahmen wurden mit Soundfield Mikrofonen durchgeführt. Somit werden im Hörversuch vier sehr unterschiedliche Aufnahmeszenarien verglichen.

Verglichene Kombinationen von Dekorrelation und Dekodierung Abbildung **??** zeigt eine graphische Darstellung der unterschiedlichen Kombinationen der Wiedergabe von Testsignalen.

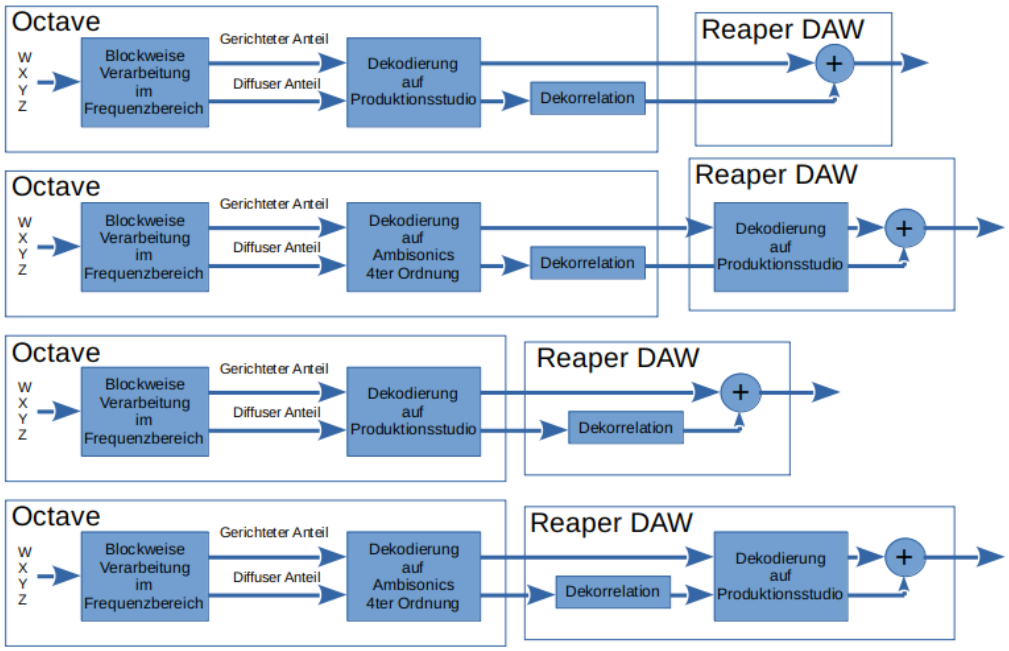

Abbildung 3 – Flussdiagram Wiedergabe im Produktionsstudio

— 12 Speaker, Random Phase Decorrelation
— 12 Speaker, FDN Decorrelation
— T-Design, FDN Decorrelation
— T-Design, Widening Plugin
— Harpex
— Compass

Wiedergabesystem Als Wiedergabesytem wurde die Lautsprecheranordnung des Produktionsstudios des IEM verwendet.

Versuchsinterface Um die Antworten der Versuchspersonen auszuwerten wurde die MUSHRA-Anwendung eingesetzt. Diese Anwendung wird über eine *.json*-Datei konfiguriert.

3.3 Ergebnis

ergebnistext

3.4 Diskussion

diskussionstext

4 Zusammenfassung und Ausblick

zusammenfassungstext

Literatur

[FZ19] M. F. Franz Zotter, *Ambisonics*. Springer, 2019.