Masterarbeit

Audioproduktion im Kontext von stereo-upmixfähigen Wiedergabesystemen

Lukas Benedicic

FH JOANNEUM University of Applied Sciences

Universität für Musik und darstellende Kunst Graz

Unter Betreuung von Martin Rieger

zur Erlangung des akademischen Grades

Master of Arts (M. A.)

Februar 2021

Danksagung

Ich möchte mich bei meinem Betreuer Martin Rieger für die Unterstützung und wertvollen Anregungen bedanken. Ein großes Danke geht auch an alle Personen, die mir im Laufe dieser Arbeit mit ihrem Wissen weitergeholfen haben:

Daniela Rieger vom Fraunhofer Institut
David Rädler von der Firma Dolby Germany
Frederick Rickmann von der Firma Steenssen
Helge Schwarz vom Bayerischen Rundfunk
Sebastian Kraft von der Helmut Schmidt Universität
Johannes Kares von der Firma Sennheiser
Jordi Alba von der Firma Timber3D

Weiters möchte ich mich bei meinen Studienkolleg*innen für den angenehmen und meist lustigen Studienverlauf der letzten Jahre bedanken, bei meiner Familie für die permanente Unterstützung, und last, but not least, bei allen Menschen die mich bewusst oder unbewusst mental unterstützt oder inspiriert haben.

Eidesstattliche Erklärung

Ich erkläre ehrenwörtlich, dass ich die vorliegende Masterarbeit selbstständig angefertigt und die mit ihr verbundenen Tätigkeiten selbst erbracht habe. Ich erkläre weiters, dass ich keine anderen als die angegebenen Hilfsmittel benutzt habe. Alle aus gedruckten, ungedruckten oder dem Internet im Wortlaut oder im wesentlichen Inhalt übernommenen Formulierungen und Konzepte sind gemäß den Regeln für gutes wissenschaftliches Arbeiten zitiert und durch Fußnoten bzw. durch andere genaue Quellenangaben gekennzeichnet.

Die vorliegende Originalarbeit ist in dieser Form zur Erreichung eines akademischen Grades noch keiner anderen Hochschule vorgelegt worden. Diese Arbeit wurde in gedruckter und elektronischer Form abgegeben. Ich bestätige, dass der Inhalt der digitalen Version vollständig mit der gedruckten Version übereinstimmt. Ich bin mir bewusst, dass eine falsche Erklärung rechtliche Folgen haben kann.

Gratwein-Straßengel, am 13.01.2021

Lukas Benedicic

Kurzfassung / Abstract

Folgende Arbeit beschäftigt sich mit der Produktion von Audiocontent, der optimiert ist für die Wiedergabe über den Echo Studio Smartspeaker, als Fallbeispiel für stereo-upmix-fähige Wiedergabesysteme. Das Thema 3D-Audio wird innerhalb der Audiocommunity als eines der vielversprechendsten angesehen, doch die Etablierung in den Mainstream erscheint langwierig. Diese Trägheit könnte auf die gegenwärtig noch recht eingeschränkte Zugänglichkeit von 3D-Audio zurückzuführen sein, da diese Inhalte meist nur über spezielle Systeme wiedergegeben werden können. Die große Mehrheit der Menschen konsumieren Audiocontent nach wie vor in Stereo. Um dennoch immersive Hörerlebnisse bieten zu können, verfügen Wiedergabesysteme wie der Amazon Echo Studio oder Soundbars etwa von Sennheiser über Stereo-Upmix-Technologien, was bedeutet, dass keine speziellen Audioformate benötigt werden. Ziel der Arbeit ist es herauszufinden wie diese Technologien aus produktionstechnischer Sicht bestmöglich zu nutzen sind. Dadurch kann in weiterer Folge eruiert werden, ob bei zukünftigen Produktionen auf solche Technologien Rücksicht genommen werden sollte.

The following work deals with the production of audio content optimized for playback via the Echo Studio Smartspeaker as a case study for stereo-upmix-capable playback systems. The topic of 3D audio is considered one of the most promising within the audio community, but its establishment into the mainstream seems to be a long way off. This 'inertia' could be due to the currently still rather limited accessibility of 3D audio, as this content can usually only be played back via special systems. The vast majority of people still consume audio content in stereo. In order to still be able to offer immersive listening experiences, playback systems such as the Amazon Echo Studio or Soundbars such as those from Sennheiser use stereo upmix technologies, which means that no special audio formats are required. The goal of this work is to find out how to best use these technologies from a production point of view. This will help to determine whether such technologies should be considered in future productions.

Inhaltsverzeichnis

Abkürzungsverzeichnis

ASW	Apparent Source Width
ADSR	Attack Decay Sustain Release
CCC	Cross-Correlation Coefficient
CTC	Crosstalk Cancellation
FOA	First Order Ambisonics
HOA	Higher Order Ambisonics
HRTF	Head Related Transfer Functions
IACC	Interaural Cross-Correlation
IC	Interaural Coherence
ICLD	Interchannel Level Difference
ICTD	Interchannel Time Difference
ILD	Interaural Level Difference
ITD	Interaural Time Difference
LCR	Left-Center-Right Loudspeaker
LEV	Listener Envelopment
MS-Stereo	Mid-Side Stereo
VBAP	Vektor-Basiertes Amplitudenpanning

Einleitung

“Räumlich” und “immersiv” - zwei aktuell häufig genutzte Schlagwörter im Audiobereich. Damit werden meist Audioinhalte beschrieben, die den/die Hörer*in einhüllen, also ein Gefühl des Eintauchens oder des unmittelbar anwesend Seins vermitteln. Die Rede ist auch von 3D-Audio, ein Überbegriff für unterschiedliche Audioformate, die neben der planaren Ebene auch die horizontale abbilden können und demnach für ein dreidimensionales Hörerlebnis sorgen. In diesem mittlerweile breiten Gebiet hat sich in den letzten Jahren einiges getan, sowohl im Hinblick auf die Entwicklung neuer Technologien, als auch deren Markteinführung.
Trotzdem, die Etablierung von 3D-Audio in den Mainstream hat noch einige Hürden zu überwinden, denn diese Formate brauchen bestimmte Wiedergabesysteme oder eine unpraktikable Anzahl von Lautsprechern für eine akkurate Darstellung. Desweiteren gibt es erst eine vergleichsweise geringe Auswahl an 3D-Audiocontent, was die Attraktivität dieser Technologien vermutlich für viele verringert. Es liegt hier so etwas wie das Henne-Ei-Problem vor: Ohne Audioinhalte werden es neue Wiedergabetechnologien und Formate schwierig haben sich zu etablieren, und ohne diese Technologien nützt entsprechender Content nichts.
Im Hinblick auf Wiedergabesysteme hat sich in letzter Zeit vor allem im Bereich der Soundbars und Smartspeaker einiges getan. Diese ohnehin schon sehr marktstarken Lautsprechersysteme verfügen mittlerweile häufig über raumerweiternde Technologoien, wie Surround Sound, 3D-Audio und Stereo-Upmixfunktionen. Vorliegende Arbeit beschäftigt sich mit der Audioproduktion im Kontext von stereo-upmixfähigen Wiedergabesystemen.
Wie schon angedeutet, gibt es bezüglich 3D-Audioformaten in Sachen Verfügbarkeit und Formatkompatibilität noch die ein oder andere Hürde zu bewältigen. Stereo ist dagegen nach wie vor das meistgenutzte Audioformat und demnach fast überall verfügbar und abspielbar. Für die Wiedergabe über Soundbars und Smartspeaker die über Upmix-Technologien verfügen, kann jegliche Art von Stereo-Inhalten abgespielt werden. Das heißt, diesbezüglich fällt das Problem von mangelndem Audio-Content weg. Ziel der vorliegenden Arbeit ist es herauszufinden wie Stereo-Content produziert werden muss, damit dieser im Upmix-Kontext im Sinne einer immersiven Hörerfahrung gut funktioniert. Dadurch kann in weiterer Folge Aufschluss über das Potential von Upmixtechnologien entsprechender Wiedergabegeräte gegeben werden. Damit einhergehend kann beurteilt werden inwiefern die zu untersuchenden Produktionseigenschaften zukünftig generell in Stereoproduktionen beachtet werden sollten.
Diese Arbeit konzentriert sich auf das Thema anhand des Amazon Echo Studio Smartspeakers als Fallbeispiel für stereo-upmixfähige Wiedergabesysteme. Jener

Lautsprecher ist auf dem Gebiet besagter Geräte Vorreiter und der zurzeit wohl marktstärkste Vertreter.

Zur Struktur der Arbeit: In Hinsicht auf die Auseinandersetzung mit Audioinhalten bei denen Räumlichkeit eine zentrale Rolle spielt, wird sich einführend mit dem Thema der Hörwahrnehmung befasst. Diesbezüglich wird zunächst die Funktionsweise der menschlichen Schall-Lokalisation erklärt, damit verknüpft ist der Begriff HRTF, Head Related Transfer Functions, welcher anschließend ebenfalls näher beleuchtet wird. Ebenso wird das für die räumliche Klangwahrnehmung wichtige Konzept der Korrelation bzw. Dekorrelation von Audiosignalen beschrieben. Im Fokus steht dabei die Auswirkung von dekorrelierten Signalen auf unsere Wahrnehmung. Das Kapitel der Hörwahrnehmung wird mit der Einführung der beiden Begriffe Apparent Source Width (ASW) und Listener Envelopment (LEV) abgeschlossen. Die beiden Begriffe kommen aus der Raumakustik und werden meist zur Beschreibung der räumlichen Wirkungsweise von Konzertsälen genutzt. Jedoch erscheinen die beiden Termini auch in Bezug auf das zugrundeliegende Thema dieser Arbeit als nützlich.
Das nächste Kapitel setzt sich mit immersive Audio im Allgemeinen auseinander und bietet einen Überblick über die aktuell wichtigsten Format-Typen.
Anschließend wird auf Soundbars und Smartspeaker und deren Funktionsweise hinsichtlich der Wiedergabe von 3D-Audio eingegangen. Zu diesen gehört das sogenannte Beamforming und die Crosstalk Cancellation.
Im darauffolgenden Kapitel wird die Funktionsweise der meisten Upmix-Technologien beleuchtet. Hierbei wird vor allem auf die zugrundeliegenden Prinzipien dieser Technologien eingegangen, nämlich die Trennung von direkten und diffusen Signalanteilen. Weiters werden diverse Problematiken des Themengebiets angesprochen, dies betrifft gewisse Probleme bei der technischen Umsetzung und daraus resultierende Fehleranfälligkeiten, als auch grundlegende Definitionsschwierigkeiten.
Es folgt der Kern der Arbeit, nämlich das Analyse-Kapitel, welches sich mit der Funktionsweise der Upmix-Technologie des Echo Studio Smartspeakers auseinandersetzt. Hierfür wurden die vier gerichteten Lautsprecher mikrofoniert und folgendermaßen untersucht: Zunächst wurde das Aktivitätsmuster der Lautsprecher ermittelt, sowohl bei aktivierter Raumklangfunktion, als auch bei deaktivierter. Dazu wurden Aufnahmen von rosa Rauschen und ausgewählten Musikstücken gemacht. Diese wurden auf Lautstärke, Frequenzgang und Korrelation untersucht. Bei den Musikbeispielen stand der Fokus auf der Instrumentierung und Effektierung, sowie auf die Korrelationswerte, sowohl vom Song-File, als auch von der Wiedergabe über den Smartspeaker. So konnten erste Erkenntnisse darüber gewonnen werden, welche Produktionseigenschaften im Kontext der Upmix-Funktion besser oder schlechter funktionieren.
Anschließend wurde eine Analyse der tatsächlichen Hörerfahrung durchgeführt. Dafür wurden Binauralaufnahmen von der Abhörposition gemacht, während rosa

Rauschen in mono und in stereo, sowie die sieben Musikstücke abgespielt wurden. Diese Aufnahmen wurden bei aktivierter und deaktivierter Raumklangfunktion des Echo Studios, als auch bei der Wiedergabe über normale Stereo-Studiomonitore gemacht. Die Binauralaufnahmen dieser drei unterschiedlichen Wiedergabearten konnten dann verglichen werden um objektive Aussagen über die Wirkung der Upmix-Technologie des Smartspeakers auf den/die Hörer*in treffen zu können. Am Ende des Kapitels wurden die Analyseergebnisse diskutiert, worauf schlussendlich vier wichtige Punkte zur Audioproduktion im Kontext der Upmix-Funktion formuliert werden konnten.

In weiterer Folge wurde ein Stück produziert das einerseits die aus der Analyse gewonnen Erkenntnisse berücksichtigt, und andererseits diese Erkenntnisse prüft und erweitert. Ziel dieses Stückes ist es die Upmix-Funktion so gut wie möglich hörbar zu machen, sowie einen gewissen kompositorischen Abwechslungsreichtum zu bieten. So wurde das Stück in sechs unterschiedliche Formteile unterteilt, wobei entsprechend andere kompositorische Schwerpunkte galten. So konnten weitere Erkenntnisse zur Produktion im kontext von upmixfähigen Geräten gewonnen werden. Im Kapitel “Produktion” wird das “Room Switch” genannte Stück chronologisch nach den Formteilen A bis F in Sachen Produktion aufgeschlüsselt.

Im letzten Kapitel werden die im Zuge der Audioproduktion gewonnenen Erkenntnisse diskutiert. Dazu gehört neben den aus der Analyse formulierten Punkten, der Einfluss von kompositorischen Eigenschaften. Ebenso wird die Frage nach gewissen Mixing-Konventionen gestellt, da es zu überlegen gilt, ob Audioproduktionen im Kontext von stereo-upmixfähigen Wiedergabesystemen universell als Stereoproduktionen funktionieren sollen, oder ob es Sinn macht speziell für diese Technologie zu produzieren. Es folgt eine Einschätzung des Potentials der gewonnen Erkenntnisse aus dieser Arbeit hinsichtlich zukünftiger Audioproduktionen. Zum Schluss gibt es einen kleinen Ausblick für die mögliche weiterführende Forschung zum Thema.

1. Hörwahrnehmung

1.1 Schall-Lokalisation

Räumliches Hören bezieht sich auf die Fähigkeit unseres Gehörs, Schallquellen in einem dreidimensionalen Raum orten zu können.[1] Dazu gehört die Richtungslokalisation, als auch die Wahrnehmung der Entfernung von Schallereignissen.

Die Klangquellen-Lokalisation auf horizontaler Ebene funktioniert über Laufzeit- und Pegelunterschiede zwischen den beiden Ohren. Laufzeitunterschiede (interaural time difference, ITD) ergeben sich aus der zeitlichen Differenz die der Schall braucht um auf beide Ohren zu treffen. Der Schall wird von jener Richtung wahrgenommen in der das ipsilaterale, also das der Schallquelle zugewandte Ohr liegt.[2] Durch Laufzeitunterschiede kommt es zu Phasenverschiebungen die wir für die Lokalisation von Schall mit Frequenzen bis ungefähr 1000 Hz nutzen.[3] Die Lokalisation durch Phasenverschiebungen ist vor allem bei der Wahrnehmung von statischen niederfrequenten Schallquellen relevant.

Ebenfalls auf horizontaler Ebene funktioniert das Konzept der Pegeldifferenz (interaural level difference, ILD), diese bezeichnet den Pegelunterschied zwischen ipsilateralem und kontralateralem[4] Ohr. Die Schallquelle wird in jener Richtung wahrgenommen in der das ipsilaterale Ohr liegt, wobei die Differenz den Winkel bestimmt. Pegelunterschiede werden nur bei Frequenzen deren Wellenlänge kleiner als der Durchmesser des Kopfes ist wahrgenommen, da sich längere Wellen um den Kopf beugen und dadurch keine wesentliche Lautstärkedifferenz entsteht; die Beugung von (Schall)-Wellen wird auch Diffraktion genannt. Die Lokalisation durch Pegeldifferenzen funktioniert bei nahegelegenen Quellen auch für tiefere Frequenzen.[5]

Nur über Laufzeit- und Pegelunterschiede ist es jedoch nicht möglich zu bestimmen ob ein Klang von vorne oder von hinten kommt, da sich diese Differenzen ausschließlich auf den lateralen Bereich der Lokalisation beziehen. Diesbezüglich hätte man bei einer Schallquelle die horizontal gesehen bei 45° erklingt, also von rechts vorne kommt, die gleiche Hörerfahrung wie bei einer Schallquelle die von

[1] vgl. Wenzel, Begault, Godfroy-Cooper (2017), S.10
[2] vgl. Wenzel, Begault, Godfroy-Cooper (2017), S.11
[3] vgl. Wenzel, Begault, Godfroy-Cooper (2017), S.13
[4] Ipsilateral bedeutet “auf der gleichen Seite des Körpers”, kontralateral bedeutet “auf der entgegengesetzten Seite des Körpers”. Im Kontext bezeichnet ‘ipsilateral’ also das Ohr das näher an der Schallquelle ist und ‘kontralateral’ demnach das entgegengesetzte.
[5] vgl. Wenzel, Begault, Godfroy-Cooper (2017), S.14

135° kommt, also von rechts hinten. Dies ist der Fall, weil die Laufzeit- und Pegeldifferenz von beiden genannten Winkel die gleiche wäre. Dasselbe Prinzip gilt für die Klanglokalisation auf vertikaler Ebene, wodurch auch diese nicht mit Laufzeit- und Pegelunterschieden zu erklären ist. Daher sind wir auf die richtungsabhängigen Filterungen angewiesen, die durch Reflektionen die aufgrund der Form unseres Außenohres, als auch durch andere physische Gegebenheiten wie der Form der Schultern und des Torsos entstehen.[6]

Nicht zu vernachlässigen sind die akustischen Eigenschaften des Raumes in dem man sich befindet, diese nehmen aufgrund der Reflektionen starken Einfluss auf die räumliche Hörwahrnehmung.[7] Reflektionen geben uns durch ihr Verhältnis zum Direktschall Aufschluss zur Distanz der Schallquelle.[8] Ist der zeitliche Abstand von Direktschall und Reflektionen vergleichsweise groß, wird die Schallquelle als nahe empfunden. Zu den Reflektionen sind in Bezug auf die Distanzwahrnehmung die Lautstärke, das Timbre, der zeitliche Verlauf des Klangs (ADSR-Kurve), sowie die kognitive Vertrautheit relevant.[9]

Unterschiedliche Umgebungseigenschaften lassen uns also Schallquellen verschieden wahrnehmen und lassen somit auch Rückschlüsse auf die Umgebung selbst zu. Diesbezüglich sind Reflektionszeit und das Verhältnis zwischen den ersten Reflektionen und Direktschall, aber auch die Art der räumlichen Verteilung der frühen Reflektionen wichtig. Die Umgebung liefert zudem Hintergrundgeräusche die uns bei der Kontextualisierung von Schallquellen unterstützt. Studien legen zudem nahe, dass wir uns an die Reflektionsmuster von Umgebungen anpassen, sodass uns die Lokalisation von Klängen mit der Zeit leichter fällt.[10]

Die späten Reflektionen geben ebenfalls Informationen über die Umgebung. Das Auslaufen der späten Reflektionen, auch Nachhall genannt, lassen Rückschlüsse auf die Beschaffenheit der Umgebung zu. So klingt beispielsweise der Nachhall in Räumen mit vielen nicht-absorbierenden aber stark reflektierenden Oberflächen (z.B. Badezimmer) länger und heller als in jenen die mit absorbierenden Materialien eingerichtet sind (z.B. Wohnzimmer).[11]

Dass Raumreflektionen für unsere Klangwahrnehmung wichtig sind, zeigt zudem die Tatsache, dass es im Freien, was für gewöhnlich einem reflektionsarmen Umfeld entspricht, eine erhöhte Fehlerquote bei der Ortung von statischen Sounds gibt. Grundsätzlich schwerer tun wir uns bei der Schallortung von vorne/hinten sowie oben/unten, hier kommt es häufig zu Verwechslungen.[12] Die Möglichkeit unseren Kopf zu bewegen reduziert diese „Fehleranfälligkeit“ jedoch signifikant.[13]

---

[6] vgl. Wenzel, Begault, Godfroy-Cooper (2017), S.16
[7] vgl. Wenzel, Begault, Godfroy-Cooper (2017), S.19
[8] vgl. Wenzel, Begault, Godfroy-Cooper (2017), S.27
[9] vgl. Wenzel, Begault, Godfroy-Cooper (2017), S.24
[10] vgl. Wenzel, Begault, Godfroy-Cooper (2017), S.30
[11] vgl. Wenzel, Begault, Godfroy-Cooper (2017), S.31
[12] vgl. Wenzel, Begault, Godfroy-Cooper (2017), S.17
[13] vgl. Wenzel, Begault, Godfroy-Cooper (2017), S.18

Der Grad der Lokalisierbarkeit ist auch abhängig von der Struktur des Klangs, so sind reine Töne (Sinusschwingung) deutlich schwieriger auszumachen als breitbandige Sounds. Die Lokalisierbarkeit von Klängen verbessert sich demnach mit höherer spektraler Dichte.[14] Impulsartige Sounds sind am einfachsten zu lokalisieren.[15] Wie schon angedeutet spielt auch die Vertrautheit eine Rolle, so sind gewohnte Klänge leichter zu orten, da wir ein gewisses Vorwissen über deren Spektrum haben und somit wissen inwiefern der Klang gefiltert wurde.[16] Eine weitere Begünstigung der Klanglokalisierbarkeit ist die Sichtbarkeit der Schallquelle.[17]

Im Kontext der Beschreibung von Schallquellen-Lokalisation werden die Begriffe Azimuth und Elevation benutzt. Der Azimuth beschreibt die horizontale Ebene und die Elevation beschreibt die vertikale Ebene, diese werden in Grad angegeben und gehen von -180° bis 180°. Dabei gilt für den Azimuth: Rechts vom/von der Hörer*in sind die positiven Werte, links die negativen. Bei der Elevation sind oberhalb der Hörposition die positiven und unterhalb die negativen Werte. Die Entfernung der Schallquelle wird über den Radius r angegeben.[18]

1.2 Head Related Transfer Functions

Die vorhin beschriebenen Mechanismen zur Klanglokalisation lassen sich letztendlich als Filterkurven beschreiben, diese werden Head Related Transfer Functions genannt, oder kurz HRTF.

HRTFs werden benutzt um räumliche Audioszenen als binaurales Signal über Kopfhörer abzubilden. Zugrunde liegt der Gedanke, dass es unter der Voraussetzung einer adäquaten Klangformung des Audiosignals nach Vorbild unserer Lokalisierfähigkeit möglich ist, Audioszenen realitätsgetreu darzustellen.[19]
Diese Filterkurven können mithilfe von im Ohrkanal platzierten Mikrofonen oder sogenannten Kunstkopfmikrofonen gemessen werden. Durch diese Platzierung, ob in echten Ohren oder wie im Falle eines Kunstkopfmikrofons, nachgebildeten Ohren, wird eine natürliche räumliche Hörerfahrung eingefangen. Die Wiedergabe dieser Aufnahmen ist in der Regel für Kopfhörer bestimmt, da es zu keinem Übersprechungseffekt zwischen den beiden Ohren kommen sollte, welcher die akkurate akustische Raumdarstellung verzerren würde. Die durch dieses Verfahren aufgenommenen und wiedergegebenen Audiosignale nennt man binaurales Audio.
Das unvorteilhafte dieser Technologie, also der binauralen Darstellung von Klang, ist jedoch, dass HRTFs individuell verschieden sind. Zwar funktionieren die meisten

[14] vgl. Käsbach (2016), S. 10
[15] vgl. Wenzel, Begault, Godfroy-Cooper (2017), S.18
[16] vgl. Wenzel, Begault, Godfroy-Cooper (2017), S.19
[17] vgl. Wenzel, Begault, Godfroy-Cooper (2017), S.23
[18] vgl. Wenzel, Begault, Godfroy-Cooper (2017), S.11
[19] vgl. Wenzel, Begault, Godfroy-Cooper (2017), S.19

HRTFs für den Großteil der Menschen einigermaßen gut, wirklich perfekt funktionieren allerdings nur individualisierte HRTFs.

1.3 Dekorrelation

Hinsichtlich der räumlichen Klangwahrnehmung ist das Konzept der Korrelation und Dekorrelation integraler Bestandteil. Das gilt sowohl für die Audioreproduktion als auch für natürliche Umgebungen.[20] Dekorrealtion ist maßgeblich für die empfundene Breite eines Signals verantwortlich, beispielsweise bei der Stereowiedergabe.
Der Korrelationsgrad wird von -1 bis +1 angegeben, wobei +1 bedeutet, dass die Signale ident und gleichphasig sind, also korrelieren. Ein Wert von -1 bedeutet, dass die Signale ident, jedoch um 180° phasenverschoben sind, und 0 entspricht unterschiedlichen Signalen, hier spricht man von unkorrelierten Signalen. Der Wert wird Kreuz-Korrelations-Koeffizient (Cross-Correlation Coefficient, CCC) genannt.[21] Man spricht allgemein von interauraler Kreuz-Korrelation.

Kendall nennt fünf Effekte, die Dekorrelation auf unsere räumliche Hörwahrnehmng hat:

1. Timbrale Färbungen und Kammfilter sind nicht mehr wahrnehmbar.
2. Dekorrelierte Kanäle produzieren diffuse Schallfelder ohne Reflektionen.
3. Dekorrelierte Kanäle erzeugen externe Hörerlebnisse über Kopfhörer (Externalisation).
4. Es kommt zu keiner wahrnehmbaren Positionsverschiebung des Klangbildes (image shift, siehe ad 4) beim Wechsel der Hörposition.
5. Das Problem des Prezedänz-Effekts, welcher das Stereobild jeweils in den nächsten Lautsprecher kollabieren lässt, wird verhindert.[22]

Ad 1. Timbrale Färbungen und Kammfilter entstehen durch konstruktive und destruktive Interferenzen. Es wurde festgestellt, dass diese Effekte durch Dekorrelation eliminiert werden können, dies ist der Fall bei einem Koeffizienten von (nahe) 0. Eliminiert werden die wahrnehmbaren Effekte der Interferenzen, physikalisch sind sie noch vorhanden.[23]

Ad 2. Die wahrgenommene Qualität von diffusen Signalen ist abhängig von der interauralen Kreuz-Korrelation, also von der Unterschiedlichkeit der Signale die auf unsere Ohren treffen. Hören wir etwas von vorne, treffen die Schallwellen auf beide Ohren gleichzeitig und sind daher korreliert (obwohl nicht 100%ig). Hallfahnen

[20] vgl. Kendall (1995), S.71
[21] vgl. Kendall (1995), S.72
[22] vgl. Kendall (1995), S.71
[23] vgl. Kendall (1995), S.78

treffen unkorreliert auf unsere Ohren und haben daher einen Koeffizienten von etwa 0, dennoch können wir den Klang der Quelle, also dem Direktsignal, zuordnen.[24]

Das breiteste Stereobild erhält man, wenn der Cross-Correlation Coefficient bei etwa 0 liegt. Der Effekt des CCC ist für Frequenzen speziell unter 1 KHz größer als für die Höhen ab 3 Hz.[25]

Ad 3. Dekorrelation ist ein essentieller Bestandteil von Hall und ist auch entscheidend für die Externalisation beim Hören mit Kopfhörern. Zudem scheint Dekorrelation auch bei der Externalisation von Direktsignalen hilfreich zu sein, wenn diese als unkorrelierte Hallfahne beigefügt wird, wobei der Effekt abhängig von der Beschaffenheit des Signals ist.[26]

Ad 4. Bei einem Delay zweier identischer Signale von unter 1 ms, wiedergegeben von zwei Lautsprechern, werden diese als ein Signal in der Mitte der beiden Lautsprecher wahrgenommen. Dieser Effekt wird „image-shift“ genannt. Dekorrelierte Signale sind gegen den „image shift“ nahezu immun.[27]

Ad 5. Der Präzedenz-Effekt wird auch Gesetz der ersten Wellenfront genannt. Er beschreibt die Lokalisierbarkeit des Direktschalls unabhängig der Reflektionen, dies betrifft vor allem transiente Klänge. Bei einer Abhörposition die außerhalb des Sweet Spots liegt, führt dies dazu, dass das Stereobild in den nähergelegenen Lautsprecher kollabiert, da sich die Ankunftszeiten der Signale zueinander verändern. Bei dekorrelierten Signalen ist es möglich das Signal über mehrere Lautsprecher gleichzeitig wahrzunehmen.[28]

1.4 Apparent Source Width und Listener Envelopment

Klangquellen werden in der Regel nicht als punktuelle Ereignisse wahrgenommen, da vor allem durch Raumreflektionen aber auch durch unseren eigenen Torso die wahrgenommene Klangbreite variiert. Diese Klangbreite wird auch Apparent Source Width (ASW) genannt.[29] Als Listener Envelopment (LEV) wird der Grad der akustischen Einhüllung der/des Hörers*in bezeichnet. ASW und LEV drücken gemeinsam das räumliche Klangerleben aus.[30] Laut Studien von Berg und Rumsey (2001) und Bech (1998), geht hervor, dass eine größere Klangbreite von Hörer*innen bevorzugt wird.[31]

[24] vgl. Kendall (1995), S.79
[25] vgl. Kendall (1995), S.79
[26] vgl. Kendall (1995), S.80
[27] vgl. Kendall (1995), S.81
[28] vgl. Kendall (1995), S.82
[29] vgl. Käsbach (2016), S.8
[30] vgl. Pfanzagl-Cardone (2010), S.31
[31] vgl. Pfanzagl-Cardone (2010), S.42

Die beiden genannten Begriffe werden zur Beschreibung von räumlichen Audioerfahrungen vor allem im Kontext von Konzerthallen genutzt, wo gute Raumakustik ausschlaggebend ist.[32] Doch in Zeiten von 3D-Audio-fähigen Wiedergabesystemen bietet es sich an jene Begriffe auch auf diesen Sektor anzuwenden.

Die Wahrnehmung der Entfernung von Schallquellen hängt im Freien von der Intensität und dem Gehalt der hohen Frequenzen ab, in Räumen ist das Direktsignal-zu-Reflektionen-Verhältnis (D/R-Verhältnis) ausschlaggebend.[33] Dieses Verhältnis liegt auch der Bewertung von ASW und LEV zugrunde:
Die ASW wird durch die ersten Reflektionen, die innerhalb von ca. 80 ms nach dem Direktschall eintreffen bestimmt und das LEV wird demgegenüber von den diffuseren Reflektionen nach ca. 80 ms bestimmt.[34] Dabei hängt die Breite der ASW davon ab wie stark die eintreffenden Reflektionen korrelieren, und auch das LEV bezieht sich auf die Dichte und die räumliche Verteilung von Reflektionen die jedoch später auf die Ohren treffen.[35] Ein diffuses Schallfeld besteht in der Theorie aus unendlich vielen unkorrelierten ebenen Wellen, die aus allen Richtungen kommen. Solche Schallfelder tragen maßgeblich zu einer umhüllenden Audioerfahrung bei.[36]

Für die Wahrnehmung der ASW spielen drei binaurale Faktoren eine Rolle: Interaurale Zeitdifferenz (ITD), interaurale Pegeldifferenz (ILD), und interaurale Kohärenz (IC). Durch Raumreflektionen und Reflektionen vom Torso fluktuieren diese Faktoren auf zeitlicher Ebene. Je mehr Raumreflektionen vorhanden sind, desto kleiner wird die IC und desto variabler werden ITDs und ILDs, was eine größere ASW zur Folge hat.[37]
ASW und LEV sind zunächst subjektive Maße. Die interaurale Kreuz-Korrelation (IACC) ist dagegen ein objektives Maß, das über die Zeitkorrelation der beiden Ohren berechnet wird.[38] Damit lassen sich ASW und LEV objektiv bewerten. IACC beschreibt die Ähnlichkeit von Signalen in einem Zeitfenster von 1ms. Ein niedriger Korrelationsgrad führt zu einer breiteren Klangwahrnehmung. Zur Messung der ASW werden die ersten 80 ms nach Eintreffen des Direktschalls auf den IACC untersucht und bei der Messung vom LEV die Reflektionen nach 80 ms.[39]

Neben der interauralen Kreuz-Korrelation hängt die ASW und das LEV von weiteren Faktoren ab:

[32] vgl. Lee (2013), S.1
[33] vgl. Käsbach (2016), S.11
[34] vgl. Lee (2013), S.1
[35] vgl. Avni, Rafaely (2009), S. 1
[36] vgl. Cousins, Bleeck Fazi (2017), S. 1
[37] vgl. Käsbach (2016), S.76
[38] vgl. Avni, Rafaely (2009), S. 1
[39] vgl. Lee (2013), S.3

Die spektralen Eigenschaften eines Klangs haben ebenso Einfluss auf die ASW und das LEV.[40] Die ASW wird durch Reflektionen im Frequenzbereich zwischen 1 kHz und 10 kHz beeinflusst. Beim LEV spielen hingegen tiefere Frequenzen unterhalb von 200 Hz bis etwa 500 Hz eine tragende Rolle. Den größten Einfluss auf die ASW nehmen die Frequenzbereiche 500, 1000, und 2000 Hz.[41] Wobei es hier unter den Forschern unterschiedliche Auffassungen gibt. Käsbach beschreibt den Bereich um 1 KHz als besonders relevant für eine gesteigerte räumliche Wahrnehmung. Dieser Frequenzbereich beeinflusst die ASW-Wahrnehmung, Frequenzen über 2 kHz beeinflussen diese hingegen nicht.[42] Unabhängig vom Reflektionsmuster fanden Blauert und Lindemann (1986) heraus, dass tiefe Frequenzen bei gleichbleibendem IACC eine größere ASW haben.[43]

Eine optimale Einhüllung ist zudem abhängig vom Einfallswinkel des Schalls, dieser Winkel ist wiederum frequenzabhängig. Signale unter 700 Hz sollten beispielsweise möglichst seitlich bei dem/der Hörer*in eintreffen. Bei einem Signal mit der Frequenz von 1000 Hz das im Winkel von 45° eintrifft gibt es eine große Fluktuation der interauralen Zeitdifferenz (ITD) was ein starkes Gefühl der akustischen Einhüllung zur Folge hat. Bei einem Signal von 2000 Hz das im selben Winkel eintrifft, findet hingegen nur eine geringe Fluktuation der ITD statt (siehe Abb. 1).[44] Fluktuationen innerhalb der ITDs und ILDs führen daher zu einer geringeren Kohärenz zwischen den zwei an den Ohren eintreffenden Signalen. Man kann also sagen, dass die ASW und das LEV invers proportional zur IACC ist: Je größer die ASW bzw. das LEV, desto kleiner die IACC.[45]

[40] vgl. Sato, Ando (2001) S.1
[41] vgl. Pfanzagl-Cardone (2010), S.32
[42] vgl. Käsbach (2016), S.99/100
[43] vgl. Käsbach (2016), S.29
[44] vgl. Pfanzagl-Cardone (2010), S.34
[45] vgl. Käsbach (2016), S.11

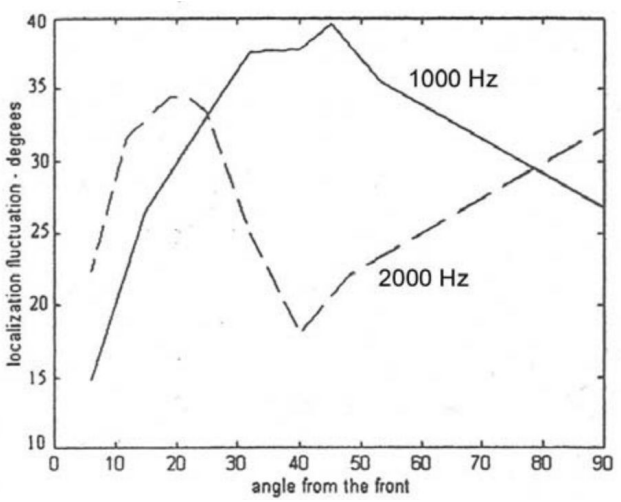


Abb. 1: Frequenz- und Winkelabhängige Unterschiede der Fluktuationen von ITDs und ILDs.

Ebenfalls ausschlaggebend für das LEV ist die Energie-Relation zwischen den Reflektionen von vorne und von hinten. Je kleiner dieses Energieverhältnis ist, desto größer ist das empfundene LEV. Es wurde allerdings noch kein Messverfahren für dieses Verhältnis vorgestellt. Lee berechnet in seiner Studie das „B/F“-Verhältnis (Back-Front) aus der Energieverteilung der virtuellen Front- und Rear-Signale aus der W- und X-Komponente einer Ambisonics B-Format Impulsantwort.[46]

Des weiteren nimmt auch der Schalldruckpegel Einfluss auf die ASW. Das heißt je lauter das Signal desto größer ist die ASW.[47]

Die ASW und das LEV, die gemeinsam das räumliche Hörerlebnis beschreiben, hängen also in erster Linie von den Reflektionen ab. Diese wiederum werden durch interaurale Zeit- und Pegeldifferenzen, sowie Kohärenzen wahrgenommen. Mit dem IACC können diese Differenzen objektiv durch Zahlenwerte ausgedrückt werden. Weiters abhängig ist die ASW und das LEV von der spektralen Beschaffenheit der Signale, dem Einfallswinkel, dem “Back-Front Energieverhältnis”, und dem Schalldruck.

[46] vgl. Lee (2013), S.1
[47] vgl. Käsbach (2016), S.35

2. Immersive Audio

Immersiver Klang umhüllt uns bestenfalls omnidirektional, es entsteht ein Gefühl des “Eintauchens” oder „anwesend Seins“.[48]

In unserem Alltag sind wir permanent von Höreindrücken aus allen Richtungen umgeben, wir befinden uns in einer natürlichen Hörumgebung. Im Gegensatz dazu steht die virtuelle Hörumgebung, welche mit Hilfe von Lautsprechern oder Kopfhörern produziert wird, mit dem Ziel unser Hörerlebnis von der natürlichen Hörumgebung zu entkoppeln oder zu erweitern.[49] Das Ziel einer immersiven virtuellen Hörumgebung ist es, so nahe wie möglich an eine natürliche Hörerfahrung, im Sinne von natürlich lokalisierbar, heranzukommen.
Die Wahrnehmung von Immersion ist das Produkt einer Vielzahl an komplexen Mechanismen und Interaktionen zwischen Menschen und Schallwellen (siehe Kapitel Hörwahrnehmung).[50] Somit sind immersive Hörerfahrungen perzeptiv multidimensionale Erlebnisse.[51]
Virtualisierungen von auditiver Räumlichkeit können demnach am Besten unter Berücksichtigung der relevanten akustischen und psychoakustischen Phänomene realisiert werden.[52]
Unabhängig von der Wiedergabeart, sei es Stereo, Surround, oder 3D-Audio, das Gefühl von Immersion wird immer abhängig von jenen akustischen und psychoakustischen Faktoren sein.[53] Um an dieser Stelle Verwirrung vorzubeugen, sei festgehalten, dass der Terminus “immersiv” nicht gleich 3D-Audio bedeutet. Auch beispielsweise normale Stereowiedergabe weist einen gewissen Immersions-Grad auf, dieser ist bei 3D-Audio in der Regel eben größer. Zudem muss der Begriff nicht zwangsläufig in einen räumlichen Kontext gesetzt werden, auch eine Geschichte kann immersiv sein. Der Begriff “Immersive Audio” wird hingegen meist synonym zu Spatial Audio oder 3D-Audio verwendet.

Audioreproduktion kann grundsätzlich kategorisiert werden in Kopfhörer- und Lautsprecherwiedergabe.[54] Wichtiger Bestandteil bei der Erzeugung von räumlichen Audiocontent ist das Rendering. Darunter versteht man die Darstellung von Klangobjekten mittels mehrerer Lautsprecher, sodass diese Objekte an der intendierten Stelle im Raum wahrgenommen werden können. Rendering-Algorithmen kommen sowohl bei kanalbasierten als auch bei

[48] vgl. Roginska, Geluso (2017), S.1
[49] vgl. Roginska, Geluso (2017), S.2
[50] vgl. Wenzel, Begault, Godfroy-Cooper (2017), S.33
[51] vgl. Wenzel, Begault, Godfroy-Cooper (2017), S.24
[52] vgl. Wenzel, Begault, Godfroy-Cooper (2017), S.34
[53] vgl. Wenzel, Begault, Godfroy-Cooper (2017), S.34
[54] vgl. He (2016), S. 2

kanalunabhängigen Wiedergabesystemen zum Einsatz.[55]
Der Vorgang der Klangreproduktion kann als Quelle-Medium-Empfänger Modell gesehen werden. Dabei ist mit Quelle der Audiocontent gemeint, Medium steht für das Wiedergabesystem, und die Empfänger sind unsere Ohren. Um natürliche immersive Klangerlebnisse schaffen zu können, muss Konsistenz zwischen den genannten Komponenten bestehen.[56]

Unsere auditive Lokalisierungsfähigkeit ist mitunter abhängig von der Anzahl der konkurrierenden Schallquellen, je mehr Schallquellen auf uns einwirken, desto schwerer tun wir uns mit der Ortung.[57] Dies spricht im Produktions-Kontext für eine Simplifizierung der immersiven Audioszene.

Im Folgenden sollen nun die aktuell wichtigsten immersiven Wiedergabekonzepte und Formate näher betrachtet werden:

2.1 Stereo und binaurales Audio

Das zweikanalige Stereoformat ist seit der Markteinführung in den 1950er Jahren das dominierende Format in Sachen Produktion und Wiedergabe.[58] Das kanalbasierte Format wird üblicherweise über zwei Lautsprecher wiedergegeben. Befindet man sich im sogenannten Sweet Spot, der optimalen Abhörposition (siehe Abb. 2), können gerichtete Klänge auf planarer Ebene zwischen den Lautsprechern akkurat dargestellt werden. Die Richtungseindrücke können entweder durch inter-kanalige Zeitunterschiede (ICTD) und/oder durch inter-kanalige Pegelunterschiede (ICLD) realisiert werden. ICLDs funktionieren aufgrund unserer Kopfgröße für höhere Frequenzen besser, da sich tiefere Frequenzen, ab etwa 300 Hz abwärts, um den Kopf beugen und und somit keine wesentliche Pegeldifferenz stattfindet. Doch da die Lautsprecher im Normalfall (vorne) seitlich vom/von der Hörer*in platziert sind, ergeben sich aus den ICLDs interaurale Zeitdifferenzen, wodurch sich die Ortung der tiefen Frequenzen für uns erleichtert.[59]

[55] vgl. Tsingos (2017), S. 247
[56] vgl. He, Gan (2015), S.1
[57] vgl. Tsingos (2017), S. 261
[58] vgl. Geluso (2017), S.63
[59] vgl. Geluso (2017), S.64

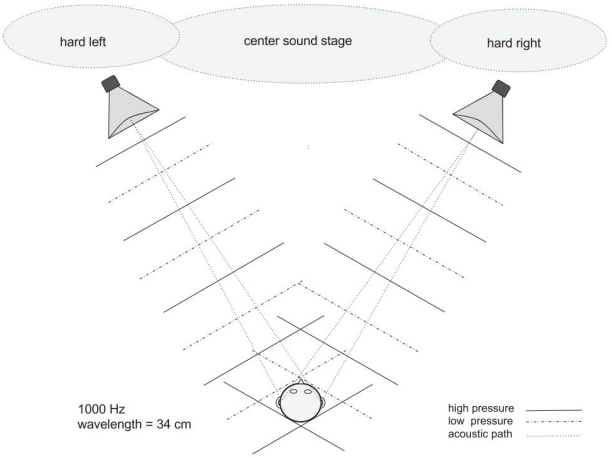


Abb. 2: Sweet Spot, Stereodreieck

Die Klangebene zwischen den Lautsprechern funktioniert aufgrund des Phänomens der Phantomschallquelle. Befindet man sich im Sweet Spot und spielt über beide Lautsprecher das gleiche Signal ab, so ergibt sich eine Phantomschallquelle in der Mitte, obwohl die beiden Schallquellen vorne seitlich vom/von der Hörer*in erzeugt werden.[60] Durch unterschiedliche Amplitudenverhältnisse lässt sich die Phantomschallquelle auf horizontaler Ebene überall zwischen den Lautsprechern platzieren.
Ebenso integral bei der Nutzung des Stereofeldes ist das Konzept der Phasenkorrelation bzw. -dekorrelation. Diese ist ausschlaggebend für die empfundene Stereobreite und Räumlichkeit. Der Korrelationsgrad wird von -1 bis +1 angegeben, wobei -1 phaseninvertierte Signale und +1 gleichphasige Signale sind. 0 bedeutet, dass die Signale dekorreliert sind, dies wirkt sich positiv auf ein breites Stereobild aus (siehe auch Kapitel Dekorrelation).[61]

In den meisten Fällen werden beim Stereoformat die beiden Kanäle als links und rechts gespeichert bzw. dargestellt, d.h. der jeweilige Kanal wird vom entsprechenden Lautsprecher wiedergegeben. Stereosignale können jedoch auch als Mittensignal plus Seitensignal betrachtet werden, auch MS-Stereo genannt. Das Mittensignal setzt sich aus den korrelierenden Teilen des Stereosignals zusammen und das Seitensignal ergibt sich aus den Phasenunterschieden der beiden Kanäle. Das Mittensignal hat also einen Korrelationswert von +1 und das Seitensignal -1. Stereosignale können immer auf beide Arten dargestellt werden, es kann auch in der Postproduktion verlustfrei gewechselt werden. Die Seitensignale beinhalten in der Praxis meist Reverb, Stereo-Delays, gepannte Quellen, etc. - also im wesentlichen

[60] vgl. Geluso (2017), S.65
[61] vgl. Geluso (2017), S.69

Rauminformationen.[62] Das MS-Stereokonzept ist mitunter im Bereich des Upmixing sehr wichtig.

Eine besondere Stereoaufnahme- und Stereowiedergabetechnik ist binaurales Audio. Binaurales Audio ist bestimmt für die Wiedergabe über Kopfhörer und kann ein 360° Klangfeld darstellen. Zugrunde liegt folgende Überlegung: Da wir sowieso nur 2 Ohren zur Verfügung haben, müsste es über Kopfhörer möglich sein eine Klangszene unter Berücksichtigung aller akustischen Faktoren realitätsgetreu wiederzugeben.[63]
Für die Aufnahme von binauralem Audio werden entweder Mikrofone die in der Ohrmuschel platziert werden verwendet, oder Kunstkopfmikrofone die dem menschlichen Hörapparat nachempfunden sind (siehe Abb. 3).

Abb. 3: Neumann KU100 Kunstkopfmikrofon

Dadurch nehmen diese Mikrofone die für den Richtungs- und Raumeindruck nötigen Filterungen, die sich durch ITDs, ILDs und den Resonanzen der Ohrmuschel ergeben, auf natürliche Weise auf. Dieses entstehende Filterungs- bzw. Frequenzprofil nennt man HRTF (siehe auch Kapitel Head Related Transfer Functions).[64]
Auf diese Art können auch Impulsantworten generiert werden, die es ermöglichen Klangszenen erst im nachhinein zu binauralisieren. Dies kommt vor allem in Kombination mit anderen 3D-Audioformaten zum Einsatz. Da man für die Wiedergabe von Formaten wie beispielsweise Ambisonics, Dolby Atmos, oder auch Surround Sound mehrere Lautsprecher benötigt, werden diese oft binauralisiert um über Kopfhörer wiedergegeben werden zu können. Dies ist vor allem bei der Produktion von Vorteil, wenn man kein entsprechendes Lautsprechersetup zur Verfügung hat.
Der Nachteil dieser Technologie ist, dass jeder menschliche Hörapparat

[62] vgl. Geluso (2017), S.68/69
[63] vgl. Roginszka (2017), S.88
[64] vgl. Choueiri (2017), S.125

unterschiedlich ist, was heißt, dass jeder Mensch ein anderes HRTF-Profil besitzt. Es gibt also kein HRTF-Profil das für alle Menschen gleich gut funktioniert. Um ein konsistentes 3D-Erlebnis zu gewährleisten, müsste jede*r Hörer*in Audioinhalte mit individualisierten HRTFs hören. Bei Kunstkopfmikrofonen und entsprechenden Render-Algorithmen setzt man daher auf durchschnittliche Kopfgrößen, sodass der 3D-Effekt für möglichst viele Personen passabel funktioniert.
Bei der Wiedergabe von binauralem Audio war bisher nur von Kopfhörern die Rede, allerdings gibt es Bestrebungen zur Lautsprecherwiedergabe, was jedoch technisch recht anspruchsvoll ist. Für eine solche Wiedergabe müsste eine Auslöschung des Übersprechungssignals am kontralateralen Ohr erfolgen. Zudem müssen die Reflektionen im Abhörraum minimiert und im besten Fall eliminiert werden (siehe auch Kapitel Crosstalk Cancellation).[65]

2.2 Surround Sound

Wiedergabesysteme mit mehr als zwei Kanälen, die in einer bestimmten Form um den/die Hörer*in auf horizontaler Ebene platziert sind, können als Surround Sound-Systeme bezeichnet werden.[66]
Das Positionieren von Klangelementen wird schon seit den 1940er-Jahren mit der Einführung des Fantasound-Systems genutzt und in weiterer Folge bei Surround-Systemen wie 5.1 und 7.1. Diese Elemente wurden und werden meist noch immer bestimmten Kanälen zugeteilt, weshalb für die akkurate Wiedergabe ein bestimmtes Lautsprechersetup nötig ist.[67] Diese Systeme werden als kanaldiskret, kanalabhänhig oder kanalbasiert bezeichnet.
Der Große Nachteil von kanalbasierten Audioformaten ist, dass der Content immer für eine bestimmte Lautsprecherkonfiguration produziert werden muss. Daher gibt es diesbezüglich standartisierte Anordnungen wie 2.0, 5.1 und 7.1.[68] Damit man als Konsument*in nicht mehr auf eine unpraktikable Lautsprecheranzahl angewiesen ist um Surround Sound zu hören, bieten einige Soundbars mittlerweile Wiedergabemöglichkeiten dieser Formate.
Stereophonie und Surround Sound können allgemein als lautsprecher- und hörerzentrierte Systeme definiert werden. Jeder der Kanäle ist für die Repräsentation der Audioszene zuständig, welche sich im Sweet Spot manifestiert.[69]

[65] vgl. Choueiri (2017), S.124
[66] vgl. Rumsey (2017), S.180
[67] vgl. Tsingos (2017), S. 244
[68] vgl. Jackson, u.a. (2017), S.1
[69] vgl. Rozenn (2017), S.276

2.3 Objektbasiertes Audio

Bei objektbasiertem Audio wird die Klangszenerie mittels Audio-Objekten dargestellt, diese sind im Normalfall Monospuren.[70] Jedem Objekt werden bei der Produktion Metadaten zugeschrieben. Diese Daten beinhalten beispielsweise Informationen über die Position und Lautstärke des Klangs. Das Bedeutet die Lautsprecheranordnung für die Wiedergabe ist bei der Produktion nicht relevant, denn die Klangszene wird bei der Wiedergabe anhand der Metadaten der Objekte unabhängig von der Lautsprecherkonstellation dargestellt.[71] Das setzt voraus, dass das Rendering erst im Wiedergabemedium stattfindet.
Vorteile von objektbasierter gegenüber von kanalbasierter Audiowiedergabe sind vor allem im Kino, wo Lautsprecher im Falle einer Surround-Wiedergabe meist gruppiert sind, deutlich zu spüren. Dazu gehören bessere timbrale Eigenschaften, genauere Punktquellen, bessere Audio/Video-Interaktion und räumliche Akkuratheit, sowie die angesprochene Unabhängigkeit von bestimmten Lautsprecherkonfigurationen.[72] Zudem bietet objektbasiertes Audio dadurch, dass das Rendering erst bei der Wiedergabe erfolgt Personalisierungs- bzw. Interaktionsmöglichkeiten. Dies ist vor allem im Broadcasting-Bereich zunehmend relevant.

Rendering-Algorithmen für objektbasiertes Audio arbeiten meist mit interauralen Zeit- und Pegeldifferenzen um Klangobjekte zu platzieren, wobei jeder der Lautsprecher mittels normalisiertem Gain-Vektor das entsprechende Signal erhält um die Klangposition darzustellen. Das Signal zur Darstellung des Klangobjekts wird also von jedem Lautsprecher produziert, jedoch zu unterschiedlichen Teilen.[73]
Beim gerichteten bzw. vektorbasierten Panning nutzt man den Vektor zwischen einem Referenzpunkt, was meistens der Sweet Spot ist, und der gewünschten Objektposition zur Platzierung. Dabei werden jene zwei, oder im Falle einer 3D-Wiedergabe drei Lautsprecher verwendet, die die Objektposition so einrahmen, dass eine Phantomschallquelle entstehen kann. Gerichtetes Panning für 3D-Lautsprechersetups wird Vektor-Basiertes Amplitudenpanning (VBAP) genannt.[74]

Eine der Herausforderungen von objektbasiertem Audio ist die erhöhte Komplexität was die Manipulation, Enkodierung und Übermittlung betrifft, da hier mit einer potentiell viel größeren Anzahl an Audio-Elementen gearbeitet wird, im Gegensatz zu kanalabhängigen Formaten.[75]

[70] vgl. Tsingos (2017), S.244
[71] vgl. Jackson, u.a. (2017), S.1
[72] vgl. Tsingos (2017), S.267
[73] vgl. Tsingos (2017), S. 248
[74] vgl. Tsingos (2017), S. 249
[75] vgl. Tsingos (2017), S. 245

Die bekanntesten objektbasierten Audioformate sind Dolby Atmos (bzw. Dolby AC-4) und MPEG-H.

2.4 Ambisonics / Klangfeld

Der Klangfeld-Ansatz stellt Audioszenen gegenüber der kanalbasierten Methoden nicht-lautsprecherzentriert dar, ist also wie objektbasiertes Audio kanalunabhängig. Dabei geht es um die Erfassung, Reproduktion und Beschreibung von Schallwellen. Bei Stereo-, Mehrkanal- und Binauralsystemen steht hingegen die Darstellung von Audio-Objekten im Fokus. Der Unterschied liegt also darin, dass beim Klangfeld-Ansatz die physikalischen Eigenschaften der Klangszene dargestellt werden und nicht „nur" das Klangerlebnis für den/die Hörer*in.[76]
Ein Klangfeld ist eine Überlagerung von Direktschall, Reflektionen und Diffusschall (etc.). Ein Klangfeld kann beispielsweise mit einer Raumimpulsantwort gemessen werden, jede Klangkomponente (z.B. Direktschall, Reflektionen) kann über bestimmte Parameter charakterisiert werden, beispielsweise Ankunftszeit, Frequenzgang oder Einfallswinkel.[77]

Das Klangfeldprinzip kommt heutzutage hauptsächlich in Form von Ambisonics zum Einsatz. Dieses Aufnahme- und Wiedergabeverfahren wurde bereits in den 1970er Jahren entwickelt.[78] Anfängliches Ziel war es ein Sourround-System zu entwickeln, das es ermöglicht eine musikalische Performance zuzüglich der räumlichen Aspekte aufzunehmen und beispielsweise für den Heimgebrauch im Wohnzimmer wieder zu reproduzieren.[79]

Für derartige Aufnahmen werden spezielle mehrkapselige Sourround-Mikrofone verwendet. Um alle Richtungsinformationen zu erhalten werden bei Mikrofonen für die erste Ambisonics-Ordnung (FOA) vier Kapseln benötigt: Eine mit omnidirektionaler Richtcharakteristik (W) und drei mit Richtcharakteristik Acht, die entsprechend dem Kartesischen Koordinatensystem in die Richtungen X, Y, und Z geortet sind.[80]

Bei Ambisonics höherer Ordnung (HOA) steigt die benötigte Kanalanzahl im Faktor $(\text{Ordnung} + 1)^2$. Die Vorteile gegenüber dem FOA-Format sind dementsprechend gegeben: Je höher die Ordnung, desto höher ist die Richtungsauflösung, die Tiefenschärfe, und desto größer ist der Sweet Spot bei der Wiedergabe.[81] In der

[76] vgl. Rozenn (2017), S.276
[77] vgl. Rozenn (2017), S.276
[78] vgl. Zotter, Frank (2019), S.5
[79] vgl. Elen (2001), S.2
[80] vgl. Zotter, Frank (2019), S.9
[81] vgl. Zotter, Frank (2019), S.19/20

Wiedergabesituation muss das Aufnahmesignal erst decodiert werden. Hier erschließt sich der Vorteil von Ambisonics gegenüber herkömmlichen Surround-Formaten: Jeder Lautsprecher erhält das gleiche Signal und gibt dieses in entsprechenden unterschiedlichen Verhältnissen wieder.[82] Dadurch lassen sich Aufnahme- und Wiedergabesituation entkoppeln, denn es ist aufgrund des Dekodierverfahrens irrelevant, ob die Kanalanzahl beim Recording mit der Lautsprecheranzahl bei der Wiedergabe übereinstimmt.[83] Weiters kann das Signal für jede Lautsprecherposition dekodiert werden, wodurch unterschiedlichste Lautsprecheranordnungen in Betracht gezogen werden können.[84]

Bis jetzt war nur von ambisonischen Aufnahmen und deren Wiedergabe die Rede, mit ambisonischen Encoder-Plugins ist es jedoch auch möglich nicht-ambisonische Signale ambisonisch wiederzugeben, diese also zu encodieren.[85] Damit lassen sich - je nach Plugin - Mono- oder Multichannel-Signale in jede gewünschte Richtung verschieben, man spricht auch von Spatialisierung - dieser Terminus gilt auch für andere 3D-Audioformate.

Zwar ist Ambisonics, wie beschrieben, sehr flexibel, jedoch braucht es wie die meisten anderen 3D-Audioformate eine recht große Lautsprecheranzahl für eine überzeugende Wiedergabe. Deswegen kommt Ambisonics zurzeit meist in binauraliserter Form zum Einsatz, im Kontext von 360°-Video oder Virtual Reality Anwendungen.

3. Soundbars und Smartspeaker

Für die Wiedergabe von 3D-Audio benötigt man entweder eine für den Heimgebrauch nicht praktikable Lautsprecheranzahl oder man muss sich auf Kopfhörer beschränken. Hersteller von 3D-Audiofähigen Soundbars und Smartspeaker sind dabei dieses Problem lösen.
Besagte Wiedergabegeräte sind mittlerweile eine der beliebtesten Wege Audio zu konsumieren. Zwar ist die Anzahl an 3D-audiofähigen Produkten noch überschaubar, doch in Anbetracht der Beliebtheit dieser Geräte, scheint es eine große Chance für die Einführung von 3D-Audio in den Mainstream zu sein.

Soundbars sind meist längliche, quaderförmige Wiedergabesysteme die häufig in Kombination mit einem Fernseher genutzt werden. In einer Soundbar sind üblicherweise mehrere Lautsprecher verbaut.

[82] vgl. Elen (2001), S.2
[83] vgl. Sontacchi, Höldrich (2000), S.6
[84] vgl. Elen (2001), S.2
[85] vgl. Zotter, Frank (2019), S.92

Smartspeaker sind Wiedergabesysteme die mit einem Voice-User-Interface verknüpft sind. Mit dem Amazon Echo Studio kam Ende 2019 ein 3D-audiofähiges Gerät auf den Markt, das wie die meisten Soundbars mehrere Lautsprecher verbaut hat.

Bei Soundbars und Smartspeaker gibt es allgemein zwei Arten um 3D-Audio wiederzugeben: Mittels gerichteten Wandreflektionen, auch Beamforming genannt, oder mittels Crosstalk Cancellation, auch Virtual Surround genannt.

3.1 Beamforming

Unter Beamforming versteht man die Abstrahlung von gebündelten Schallquellen die auf reflektierende Oberflächen gestrahlt werden, sodass Spiegelquellen entstehen, wodurch man Klänge nicht von der eigentlichen Schallquelle wahrnimmt.[86]
Die Bündelung erfolgt durch die Erzeugung von konstruktiven und destruktiven Interferenzen. Bei der Schallabstrahlung entstehen in der jeweils gewünschten Richtung konstruktive Interferenzen, während sich die Schallwellen an den restlichen Stellen destruktiv überlagern.[87] Durch diskret arbeitende Lautsprecher können mehrere Kanäle in unterschiedliche Richtungen gestrahlt werden, wodurch die Möglichkeit zur Reproduktion von Surround-Mischungen und 3D-Audio entsteht.[88]
Bei dieser Technologie besteht jedoch immer Abhängigkeit vom Raum, dessen Wände und Decke die Qualität der Reflektionen bestimmt. Weiters ist auch die Architektur relevant, wenngleich sich die meisten “beam-fähigen” Geräte mittlerweile durch Sensoren auf unsymmetrische Räume einstellen können. Zu große Entfernungen zu Reflektionsflächen bewirken, dass die räumliche Darstellungsschärfe abnimmt und der immersive Effekt verschwimmt.[89]
Für ein immersives Erlebnis eignen sich breitbandige Klänge besser als impulsartige, da man hier ein einhüllenderes Ergebnis erzielen kann. Transientenreiche Klänge tendieren dazu näher am Wiedergabegerät zu bleiben. Hörpositionswechsel führen mitunter zu starker Klangverfärbung der gerichteten Beams, da sich die Reflektionspfade ändern, was wiederum Einfluss auf die Spatialisierung haben kann.[90]

[86] vgl. Sharma (2016), S.20
[87] vgl. Sharma (2016), S.20
[88] vgl. Hooley (2006), S.355
[89] vgl. Sharma (2016), S.98
[90] vgl. Sharma (2016), S.101

3.2 Crosstalk Cancelation

Das Konzept Crosstalk Cancellation (CTC) ist nicht neu, dieses Verfahren wurde jedoch in den letzten Jahren für diverse Consumer-Devices, wie Soundbars, wiederentdeckt. Damit ist es möglich Multichannel- und 3D-Audiosignale mit einer geringen Lautsprecheranzahl wiederzugeben.[91]
Zugrunde liegt das Konzept binaurales Audio über Lautsprecher zu reproduzieren. Das heißt also, es werden Stereosignale abgespielt die mit einer HRTF encodiert werden. Da die HRTF alle Richtungsinformationen in Form von ITDs, ILDs, und entsprechenden Resonanzen enthält, setzt dies voraus, dass der rechte Kanal nur für das rechte Ohr bestimmt ist und umgekehrt - wie bei der Wiedergabe von binauralem Audio über Kopfhörer.[92]
Natürlicherweise kommt es bei Lautsprechern zu Übersprechungen, d.h. auch das contralaterale Ohr hört das Signal des entsprechenden Speakers. Um diesem Effekt entgegenzuwirken wird die Crosstalk Cancellation verwendet. Dabei wird ein zum Übersprechsignal gegenphasiges Signal abgespielt, sodass es zur Auslöschung kommt.[93]
Die Nachteile dieser Technologie sind, dass es zum Einen nur einen sehr kleinen Sweet Spot gibt, und zum Anderen, dass es durch die Auslöschungen zu recht starken timbralen Verzerrungen kommen kann.[94]
Neben der Crosstalk Cancellation sollte bei der Wiedergabe darauf geachtet werden die Raumreflektionen auf ein Minimum zu reduzieren.

[91] vgl. Baskind, u.a. (2015), S.2
[92] vgl. Choueiri (2017), S.124
[93] vgl. Sengpiel (1998), S.1
[94] vgl. Sengpiel (1998), S.1

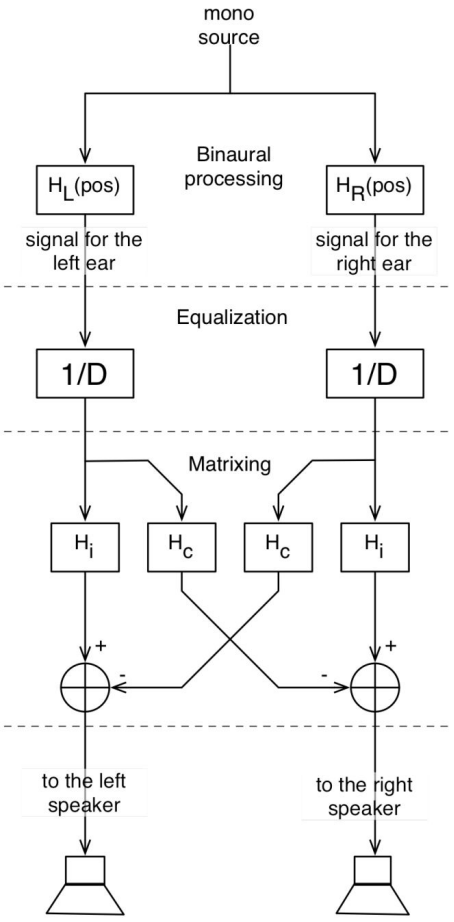


Abb. 4: Signalfluss-Konzept Crosstalk Cancellation

4. Upmixing

Unter Upmixing versteht man den Prozess Audiosignale einer niedrigeren Kanalanzahl in ein mehrkanaliges Signal umzuwandeln. Am häufigsten werden Stereosignale zu 5.1 Surround Sound geupmixed. Da kanalabhängige Formate, vor allem Stereo, nach wie vor den Großteil des produzierten Audiocontents ausmachen und es mittlerweile eine Vielzahl an unterschiedlichen Wiedergabesystemen gibt, erscheint es sinnvoll diese Formate für jegliche Systeme aufzubereiten.[95] Die Relevanz von upmixfähigen Geräten zeigt sich etwa auch im Umstand, dass große Namen wie Amazon Echo und Sennheiser neben der Wiedergabe von 3D-Audioformaten auf diese Technologie setzen.

[95] vgl. He, Gan (2015), S.1

Der Großteil der Signalverarbeitungsprozesse beim Upmixing funktioniert durch die Trennung von Mitten- und Seitensignal.[96] Das Mittensignal wird diesbezüglich meist als Primär- oder Direktsignal bezeichnet und das Seitensignal als Ambient- oder Diffussignal.
Die Wahrnehmung einer Klangszene kann grundsätzlich in Vordergrund- und Hintergrundsounds eingeteilt werden. Erstere werden demnach als direkte oder primäre Klänge bezeichnet, letztere als ambient oder diffus. Generell ist die Gliederung bzw. Trennung eines Audiosignals in die beiden Komponenten „diffus“ und „direkt“ nützlich für die Darstellung von immersiven Hörerlebnissen.[97] Aufgrund dieser unterschiedlich wahrgenommenen Komponenten, sollten diese auch unterschiedlich gerendert werden. Kanalbasierte Formate bieten jedoch nur gemischte Signale, es ist also nötig den diffusen bzw. ambient-Anteil von der primären Signalkomponente zu extrahieren.[98] Diesbezüglich werden unterschiedliche Algorithmen angewendet, die im wesentlichen jedoch den gleichen Grundgedanken verfolgen: Upmixing-Technologien unterscheiden sich dadurch wie sie den direkten vom diffusen Signalanteil trennen. Beispiele dafür sind Direct-Diffuse Decomposition oder Primary Ambient Extraction.[99]

Die wichtigste Eigenschaft zur Unterscheidung der beiden Signalanteile, ist die Korrelation zwischen den beiden Stereokanälen. Diffus-Signale haben eine schwache Korrelation und Primär-Signale haben eine starke. (Ibrahim, Allam, 2018, S. 43) Diese wird mithilfe von Inter-Kanal-Verhältnissen wie ITDs, ILDs, und dem Übersprechungskoeffizient bestimmt.[100]

Bei der Extraktion der Signale kann es durchaus zu Fehlern kommen, diese können allgemein in 3 Kategorien eingeteilt werden: 1. Verzerrungen, damit ist die unverhältnismäßige Skalierung der Amplituden des extrahierten Signals im Vergleich zum Originalsignal gemeint. 2. Interferenzen, diese werden durch die Dekorrelationsstärke des extrahierten Signals ausgedrückt 3. Verlust (Leakage), hiermit ist die Menge der unerwünschten Signalinformationen im jeweiligen extrahierten Signal gemeint, beispielsweise wenn Teile des Diffus-Signals im Primär-Signal sind (oder umgekehrt). Verzerrungen und Verluste sind stärker wahrnehmbar als Interferenzen.[101]

Stereokanäle werden meist als die Summe von Signalen plus additive unkorrelierte Signale beschrieben. Die Gewichtungen der Signale in den einzelnen Kanälen nennt

[96] vgl. Kraft, Zölzer (2015), S.2
[97] vgl. He (2016), S.3
[98] vgl. He (2016), S.2
[99] vgl. Pulkki, Delikaris-Manias, Politis (2017), S.78
[100] vgl. He (2016), S.48/49
[101] vgl. He (2016), S.47

man Panningkoeffizient, diese werden durch Werte zwischen 0 und 1 ausgedrückt.[102] Das Panning-Verhältnis sollte beim Upmixing möglichst beibehalten werden. Daher werden die Panningkoeffizienten eines jeden Sub-Bands involviert, um die Trennung von Direkt- und Ambientanteilen von Signalen die nicht direkt in die Mitte gepanned sind zu ermöglichen. Dies ist mit einer reinen Mid-Side-Trennung nicht machbar.[103]
Es ist mathematisch unmöglich die Panningkoeffizienten und Ambient-Signalanteile unabhängig voneinander zu extrahieren, in Anbetracht der 2 Kanäle als Input. Es wird jedoch angenommen, dass bei einer genug hohen Zeit- und Frequenzauflösung des Signals, an einem bestimmten Zeitpunkt und Frequenzband immer nur ein dominanter Signalanteil aktiv ist und andere Anteile vernachlässigbar klein sind. Dadurch kann die Zeit-Frequenz-Darstellung des Signals unabhängig vom Panningkoeffizienten betrachtet werden, zudem kann das Ambient-Signal simplifiziert beschrieben werden.[104]

Bei einem typischen Upmix-Szenario von Stereo zu einem Fünfkanalsystem, wird das Direkt-Signal von den Vorderen Speakern (LCR) wiedergegeben, und die rechten und linken Ambient-Signale werden auf die vier Eck-Lautsprecher L, Lr, R, Rr aufgeteilt.[105]

Wie erwähnt basieren die meisten Upmix-Methoden auf dem Prinzip der Trennung von Direkt- und Diffussignalen. Ob und wie andere Methoden funktionieren, kann in anbetracht der Tatsache, dass kommerzielle Technologien für außenstehende meist unzugänglich sind, im Kontext dieser Arbeit nicht geklärt werden.
Das Thema Upmixing ist jedenfalls allgemein ein komplexes und eines das grundlegende Definitionsschwierigkeiten aufwirft. Sebastian Kraft von der Helmut-Schmidt Universität hat mir zum Thema folgendes per Mail mitgeteilt:

“Es gibt meines Wissens kein objektives und automatisches Verfahren um den Anteil von Ambience in einem Audiosignal zuverlässig zu bewerten. Wenn man etwas darüber nachdenkt merkt man auch schnell, dass es sehr schwer ist "Ambience" für eine Schallquelle zu definieren. Ist es Breite, Nachhall, Räumlichkeit, Tiefe? Welche physikalischen Eigenschaften korrelieren damit?
[...] Insgesamt aber ein komplexes und ungelöstes Thema vor allem weil die Definition von Ambience so schwammig ist.”[106]

[102] vgl. Kraft, Zölzer (2015), S.2
[103] vgl. Kraft, Zölzer (2015), S.4
[104] vgl. Kraft, Zölzer (2015), S.2
[105] vgl. Kraft, Zölzer (2015), S.4
[106] Kraft (2020)

5. Analyse Echo Studio

In diesem Kapitel wird die Upmix-Funktion des Amazon Echo Studio Smartspeaker untersucht, welche vom Hersteller Stereo-Raumklangerweiterung genannt wird. Im Folgenden wird die Rede von Upmix- oder Raumklangfunktion sein.

5.1 Amazon Echo Studio

Der Echo Studio ist einer der ersten und sicherlich der bekannteste Smartspeaker der 3D-Audioformate unterstützt. Der Ende 2019 veröffentlichte Speaker unterstützt Dolby Atmos und MPEG-H in Form von Sonys 360 Reality Audio. Diese Formate kommen in eigens neu gemischten und gemasterten Songs zum Einsatz und können über die Prämiumversion des hauseigenen Streamingdienstes Amazon Music wiedergegeben werden. Ein immersives Hörerlebnis wird bei dem Gerät durch fünf unterschiedlich gerichtete Lautsprecher ermöglicht: 2 Mitteltöner mit 51mm Durchmesser als Seitenspeaker, 1 Mitteltöner mit 51mm Durchmesser als nach oben gerichteter Vertikalspeaker, 1 Hochtöner mit 25mm Durchmesser als Frontalspeaker und 1 nach unten gerichteter Subwoofer mit 133mm Durchmesser (siehe Abb. 5).
Durch die gerichteten Lautsprecher werden über Wandreflektionen unterschiedlich positionierte Klangquellen vorgetäuscht. Der Echo Studio passt sich dazu, laut Hersteller, mittels des eingebauten Mikrofons an die akustischen Gegebenheiten an.
Neben der Möglichkeit 3D-Audioformate wiederzugeben, bietet der Smartspeaker eine Upmix-Funktion die normale Stereoformate räumlicher erscheinen lässt. (vgl. amazon.de, 2020)

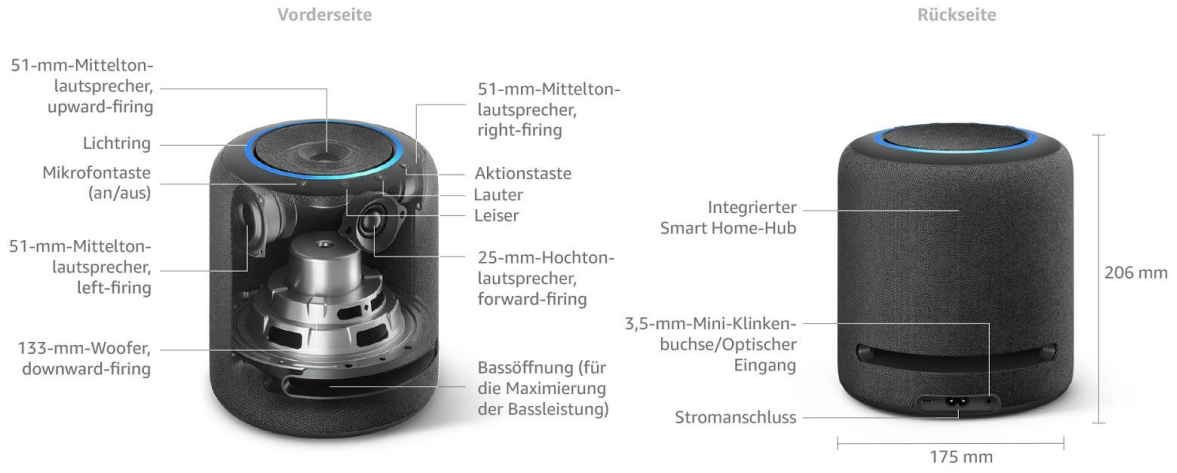


Abb. 5: Amazon Echo Studio

5.2 Analysmethode

Um herauszufinden wie die Raumklangfunktion des Echo Studios funktioniert, wurden die Seitenspeaker, der Vertikalspeaker und der Frontalspeaker mikrofoniert (siehe Abb. 6). Hierfür wurden vier Sennheiser ME64 Mikrofone verwendet. Der Subwoofer wurde nicht mikrofoniert, da dieser für die gerichteten Reflektionen nicht relevant ist. Dies hängt nicht nur an der Ausrichtung des Speakers, sondern auch an den physikalischen Gegebenheiten von tiefen Frequenzen die sich bei Lautsprechern annähernd omnidirektional ausbreiten.

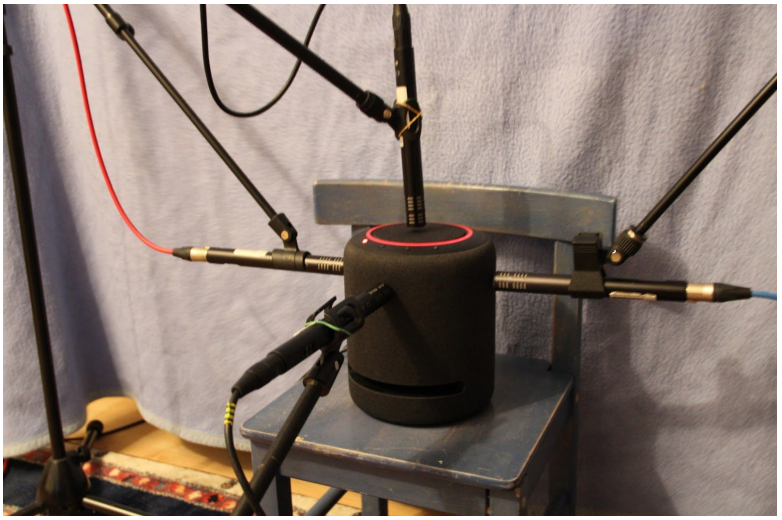


Abb.6: Echo Studio mikrophoniert

Zur Analyse der Lautsprecheraktivität wurde rosa Rauschen abgespielt. Dabei wurde das Rauschen jeweils als Monosignal, Stereosignal und als Seitensignal wiedergegeben um herauszufinden, ob und wie sich das Verhalten der Lautsprecher ändert. Die Gegenüberstellung von Mono-, Stereo-, und Seitensignal wurde als relevant erachtet, da das Gros der Upmix-Algorithmen mit einer Form der Änderung des Verhältnisses von korrelierten zu unkorrelierten Signalen arbeitet. Diesbezüglich sollte also das Korrelationsmuster der wiedergegebenen Signale Aufschluss geben.
Neben der Wiedergabe von statischen Signalen wurde auch das Panningverhalten der Lautsprecher, sowie die sukzessive Änderung des Stereobildes analysiert.
Nach der Analyse mithilfe des rosa Rauschens wurden praktische Beispiele in Form von ausgewählten Musik-Passagen unterschiedlicher Genres betrachtet.
Die wiedergegebenen Signale wurden folgendermaßen untersucht:

Zunächst wurden die Pegeldifferenzen der Lautsprecher zueinander untersucht.
Zur Ermittelung der Korrelationswerte wurde das Korrelometer-Plugin von Voxengo

benutzt. Mit diesem Multiband-Korrelometer ist es möglich das Korrelationsverhalten eines Signals über das ganze hörbare Frequenzspektrum, aufgeteilt auf bis zu 64 Bänder, zu untersuchen.
Weiters wurde das Spektrum-Analyse-Plugin “SPAN”, ebenfalls von Voxengo, benutzt. Das Frequenzspektrum wurde betrachtet, um einerseits den Frequenzbereich der Lautsprecher zu ermitteln und andererseits, um Unterschiede bei der Wiedergabe der jeweiligen Signaltypen und beim Panning zu erkennen. Da es hier frequenztechnisch um Vergleiche geht, sei an dieser Stelle erwähnt, dass der Frequenzgang der benutzten Mikrofone irrelevant ist, zumal es sich um vier idente Geräte handelt.
Für die Untersuchung der Musikbeispiele wurde zudem noch das Vektorskop-Plugin “Ozone 9 Imager” von iZotope verwendet. Damit wurde das Stereobild der Audiofiles mit dem der Seitenspeaker des Smartspeakers verglichen.

Die beschriebenen Analysemethoden wurden bei aktivierter und deaktivierter Raumklangfunktion angewendet, um eine Vergleichsbasis zu erhalten.

Anschließend wurden zur Analyse der Hörerfahrung Binauralaufnahmen gemacht. Diese wurden mit dem Korrelometer untersucht, um objektive Anhaltspunkte über die ASW und das LEV zu erhalten. Es wurde die Wiedergabe von Audiocontent über den Echo Studio mit aktiver und inaktiver Raumklangfunktion, sowie über Studiomonitore verglichen.

5.3 Signalverhalten bei aktivierter Raumklang-Funktion

Im Folgenden wird das Signalverhalten der vier gerichteten Lautsprecher des Echo Studios analysiert, während die Raumklang-Funktion aktiviert ist. Diese Funktion lässt sich in der zugehörigen Amazon Alexa-App unter den Audioeinstellungen ein- und ausschalten (siehe Abb. 7).

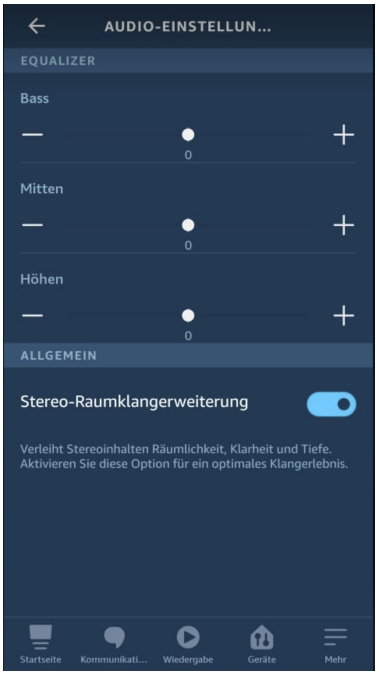


Abb. 7: Amazon Alexa-App, Audioeinstellungen

5.3.1 Monosignale

Bei der Wiedergabe von Monosignalen über den Echo Studio sind die beiden Seitenspeaker, der Frontalspeaker und der Vertikalspeaker aktiv. Am dominantesten ist hier der Frontalspeaker, welcher um 3,5 dBfs RMS lauter als die Seitenspeaker spielt. Der Vertikalspeaker ist mit -11,3 dBfs RMS im Vergleich zum Frontalspeaker nur sehr dezent aktiv.

Der Korrelationsgrad der beiden Seitenspeaker zeigt Höchstwerte bei ca. 1 kHz - 2 kHz (siehe Abb. 8). Die Werte unter 200 Hz können vernachlässigt werden, da es sich hierbei lediglich um die Übersprechung des Bassspeakers handelt; die beiden Seitenspeaker spielen erst ab ca. 200 Hz. Dennoch stellt sich die Frage warum die Korrelation ab 200 Hz nicht für alle Frequenzen den Wert 1 aufweist, da ja mono-bedingt auf beiden Seiten das gleiche Signal abgespielt werden müsste. Mögliche Erklärungen hierfür wären 1., dass die Signale durch Übersprechungen oder Reflektionen geringfügig dekorreliert wurden, oder 2., dass die Raumklangfunktion des Smartspeakers die Signale dekorreliert. Aufgrund der sehr geringen Dekorrelation kann jedoch auf eine eindeutige Erklärung verzichtet werden.

Zusammenfassend kann man sagen, dass die beiden Seitenspeaker bei der Wiedergabe von Monosignalen, wie zu erwarten, korrelieren. Zudem wird der Mono-Aspekt durch den Frontalspeaker gestützt, der das Signal mittig und nach vorne gerichtet klingen lässt.

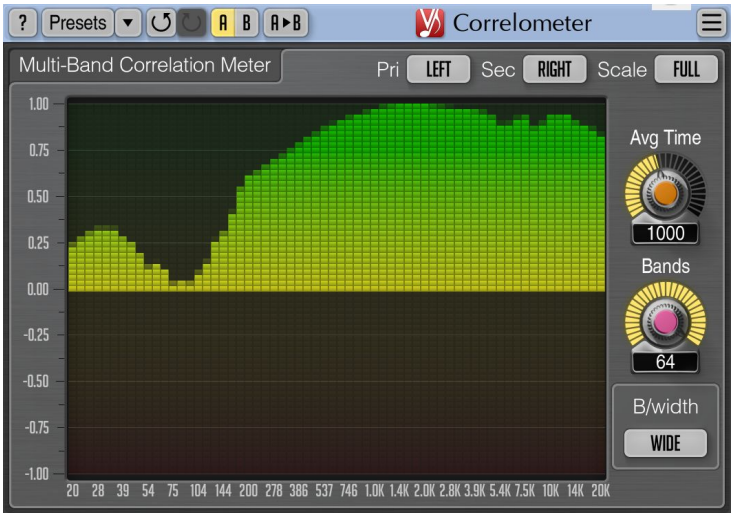


Abb. 8: Seitenspeaker, rosa Rauschen mono, Raumklang aktiviert.

5.3.2 Seitensignale

Spielt man nur das Seitensignal über den Smartspeaker ab, so sind nur die beiden Seitenspeaker aktiv.

Auffällig ist allerdings, dass die beiden Speaker nahezu perfekt korrelieren (siehe Abb. 9). Da es sich um das reine Seitensignal des Rauschens handelt, sollte jedoch das Gegenteil der Fall sein und einen Wert um -1 aufweisen. Doch wie sich herausstellt wird bei der Wiedergabe des Seitensignals bei aktivierter Raumklangfunktion die Phase einer Seite um 180° gedreht, sodass diese korrelieren.
Beim direkten Vergleich des Frequenzgangs mit dem Seitensignal bei deaktivierter Raumklangfunktion, fallen keine nennenswerten Unterschiede auf.

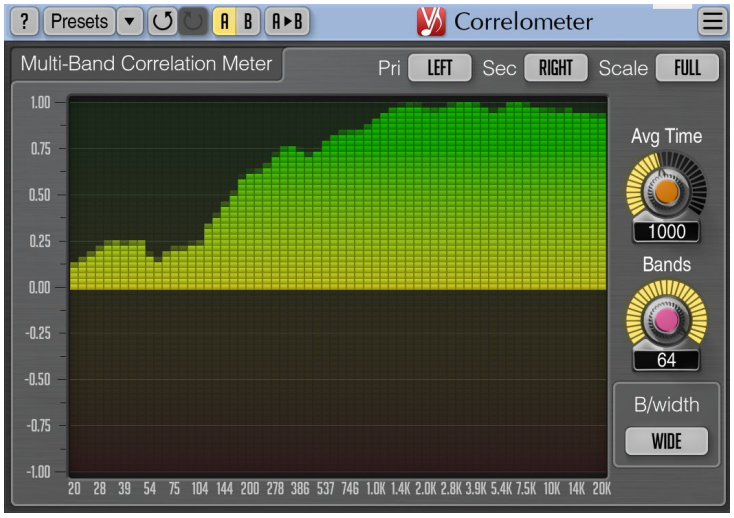


Abb.9: Seitenspeaker, rosa Rauschen Seitensignal, Raumklang aktiviert.

5.3.3 Stereosignale

Bei der ungefilterten Wiedergabe des Stereo-Files sind die Seitenspeaker, der Vertikalspeaker und der Frontalspeaker aktiv. Erstere spielen am lautesten, der Vertikalspeaker ist mit -1,2 dBfs im Vergleich zu den Seitenspeaker ebenfalls prominent, und der Frontalspeaker ist mit -15,5 dBfs sehr leise.

Die Korrelationswerte der Seitenspeaker sind bis 7,5 kHz negativ, wobei diese im Bereich um 2 kHz besonders hoch sind. Über 7,5 kHz tendiert das Signal wieder zu korrelieren (siehe Abb. 10).
Im Hinblick der Nur-Mono- und Nur-Seitensignalwiedergabe, bei der die Seitenspeaker korrelierten, stellt sich hier die Frage, wie es zu diesem Korrelationsmuster kommt.

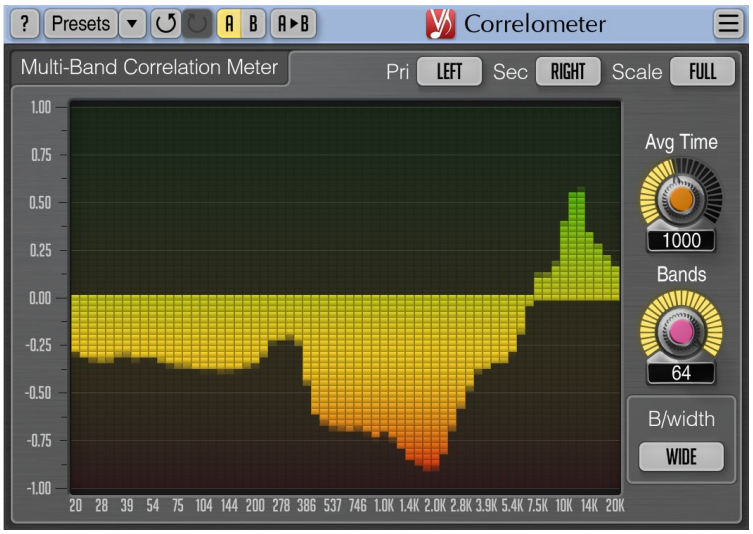


Abb. 10: Seitenspeaker, rosa Rauschen Stereo, Raumklang aktiviert.

5.3.4 Panningverhalten

Bis jetzt wurden statische Signale untersucht, doch um die Funktionsweise der Upmix-Funktion des Smartspeakers noch genauer zu eruieren, wird im Folgenden das Verhalten der Speaker beim Signalpanning analysiert.

Beim Panning des rosa Rauschens in Mono wird ersichtlich, dass hier nicht mit Amplitudenpanning gearbeitet wird, wie es bei normalen Stereowiedergabegeräten meist der Fall ist, denn selbst bei einem Hardpan in eine Richtung verändert sich die Pegelstärke der beiden Seitenspeaker nicht nennenswert. Beim Amplitudenpanning würde in diesem Fall ein Speaker inaktiv sein. Ein Blick auf den Correlometer zeigt, dass das Signal beim Panning im Bereich von 200 Hz bis 6 kHz dekorreliert. Weiters ist festzustellen, dass es zu Filterungen kommt: Beim Vergleich der beiden Seitenspeaker im Spektrum, lässt sich erkennen, dass die kontralaterale Seite eine extrem steilflankige Hight-Cut-Filterung bei 13,5 kHz aufweist, sowie einen Low-Cut bei 350 Hz. Die Höhen der ipsilateralen Seiten werden zudem angehoben (siehe Abb. 11).
Vergleicht man zudem die Seitspeaker beim Hardpan mit denen beim mittigen Pan, stellt sich heraus, dass beim Hardpan die Höhen ab etwa 4 kHz bei beiden Speaker deutlich präsenter sind als bei der mittigen Stellung (siehe Abb. 12).
Weiters nimmt die Aktivität des Frontalspeakers ab, je weiter das Signal in eine Richtung gepanned ist. Durch die Anhebung der Höhen beim Seitenpanning wird die Intensitätsabnahme des höhenlastigen Frontalspeakers womöglich kompensiert.

Das Panning bei aktivierter Raumklangfunktion arbeitet also mit einer Mischung aus Spektralpanning und phasenbasiertem Panning. Diese planare Bewegung wird durch die schwankende Aktivität des Frontalspeakers verstärkt.

Abb.11: Spektrum, rosa Rauschen mono, Hardpan, Raumklang aktiviert.

Abb.12: Spektrum, rosa Rauschen mono, mittig gepanned, Raumklang aktiviert.

5.4 Signalverhalten bei deaktivierter Raumklang-Funktion

Bei der normalen Stereowiedergabe ohne Raumklangfunktion verhält sich der Echo Studio wie ein reguläres Stereo-Wiedergabegerät: Es sind mit Ausnahme des Tieftöners, welcher immer aktiv ist, nur die beiden Seitenspeaker aktiv. Bei Monosignalen wird auf beiden Seiten das gleiche Signal wiedergegeben und bei der Wiedergabe des Seitensignals werden gegenphasige Signalanteile wiedergegeben. Das Panning funktioniert durch Amplituden-Änderung, wie man es in den meisten Fällen gewohnt ist.

5.5 Analyse von ausgewählten Musik-Passagen

Im Folgenden werden Auszüge aus unterschiedlichen Musikstücken näher betrachtet, die hinsichtlich einer immersiven Hörerfahrung über den Echo Studio gut funktionieren.
Zunächst wird die Instrumentierung ermittelt und die gehörten Sounds beschrieben. Weiters werden die benutzten Audioeffekte, wie etwa Reverb, betrachtet. Schlüsse auf die Instrumentierung und Effektierung können hier nur aus subjektiver Sicht gezogen werden, da zur Analyse lediglich die fertigen Stereo-Masterfiles und keine Stems oder gar Sessions zur Verfügung stehen.
Anschließend wird mit dem Korrelometer das Korrelationsmuster der Songpassage analysiert um die Zusammenhänge von Korrelation und der empfundenen Räumlichkeit zu überprüfen und etwaige wiederkehrende Muster zu erkennen. Es wird diesbezüglich das Audiofile, als auch die Wiedergabe der Seitenspeaker des

Echo Studios untersucht, damit Unterschiede aufgezeigt werden können. So können Rückschlüsse auf die Upmix-Funktion gezogen werden.
Abschließend wurde das Audiofile und die Wiedergabe über die Seitenspeaker mithilfe eines Vektorskops verglichen. Das Vektorskop wurde auf die Polarlevel-Darstellung eingestellt, bei der man das Energieverhältnis von Mitten- und Seitensignal, sowie die Ausbreitung im Stereofeld sieht.

5.5.1 Trentemøller - Nightwalker (0:00 - 0:16)

Zu Beginn des Stückes wird ein Gitarrenakkord angeschlagen, der mit Reverb versehen ist. Zudem hört man ein Knistern das an das einer Vinylplatte erinnert und nach ein paar Sekunden verstummt, einen Synthesizer der in den unteren Mitten angesiedelt ist, und ein Sample das dezent im Stereobild ostiniert und langsam ausfadet. Kurz bevor der Beat einsetzt erklingt eine männliche Stimme die am Ende der Phrase mit einem Ping-Pong-Delay effektiert ist.
Korrelationstechnisch hält sich dieser Introteil um den Wert 0 auf und ist somit stark dekorreliert, was sich in einem breiten Stereobild manifestiert. Der Akkordanschlag und die Stimme sind mittig im Panorama platziert und korrelieren. Der Synthesizer hingegen tendiert dazu in den negativen Korrelationsbereich auszuschlagen und ist somit relativ präsent im Seitensignal.
Frequenztechnisch gibt es zwischen 200 Hz und 300 Hz durch den Synthesizer einen dominanten Bereich und zwischen 1 kHz und 3 kHz gibt es einen weiteren ausgeprägten Bereich, welcher hauptsächlich durch den Reverb zustande kommt.
Bei der Wiedergabe über den Echo Studio mit aktivierter Raumklangfunktion zeigen die Speaker ähnliche Muster wie bei der Wiedergabe des Stereorauschens, welches ebenfalls den Korrelationswert 0 besitzt. Auffällig beim Vertikalspeaker ist, dass dieser speziell bei den hochfrequenten Hallanteilen ab ca. 6 kHz aktiv ist.
Beim Blick auf ein Vektorskop zeigt sich nochmals, dass das Stereopanorama ähnlich wie beim Stereo-Rauschen zwischen Mitten- und Seitensignal recht ausgewogen ist.

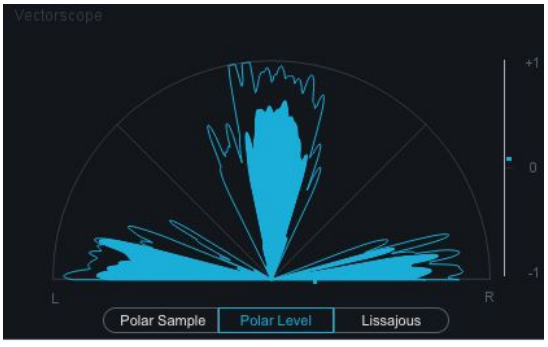


Abb. 13: Songanalyse

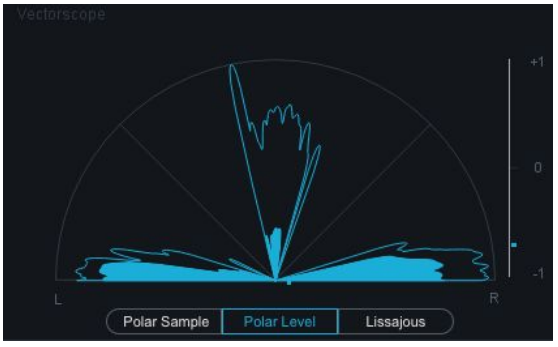


Abb. 14: Seitenspeaker

Im Vergleich zum Songfile zeigt sich bei der Vektorskop-Analyse der Seitenspeaker, dass das Mittensignal reduziert und das Seitensignal dominant ist.

5.5.2 Gabby Barrett - I Hope (0:00 - 0:20)

Das Stück beginnt mit 2 unisono spielenden E-Gitarren, wobei jede auf eine Seite gepanned ist. Dazu erklingt im Hintergrund ein nicht näher definierbares Blasinstrument. Nach 4 Takten setzen Gesang und Piano ein und nach weiteren 4 Takten kommen noch 2 im Hintergrund spielende Gitarren hinzu. Gleichzeitig setzen Fingerschnipser auf die Zählzeiten 2 und 4 ein. 4 Takte später kommt die restliche Band, also Bass und Schlagzeug, hinzu. Bei diesem Einsatz erklingt neben den Becken in den Höhen ein frequenztechnisch nach unten gleitender Soundeffekt, der entfernt an Silvesterraketen erinnert.
Die Instrumente inklusive Stimme sind mit Reverb versehen, wobei die Stimme noch mit Delay effektiert ist und klar im Vordergrund steht. Besonders gut im Sinne einer Hörereinhüllung funktionieren die Schnipser, welche mit einem Reverb versehen sind, der sich gut durchsetzt und in den Höhen sehr präsent ist. Ebenfalls gut funktioniert der Soundeffekt beim Einsatz der ganzen Band, der vor allem in den Höhen brilliert.
Die Gitarren am Anfang sind stark dekorreliert, tendenziell im negativen Korrelationsbereich. Die Stimme in der Mitte korreliert, wie zu erwarten. Die Schnipser dekorrelieren in den Höhen. Das Vektorskop zeigt ein relativ ausbalanciertes Verhältnis zwischen Mitten- und Seitensignal.
Das Korrelationsmuster der beiden Seitenspeaker des Echo Studios ist zu Beginn ab 300 Hz stark negativ, beim Einsatz des Gesangs schlagen die Korrelationswerte in den positiven Bereich. Es lässt sich tendenziell wieder das Korrelationsmuster des stereo rosa Rauschens erkennen (siehe Abb. 10), wenn auch fluktuierend.

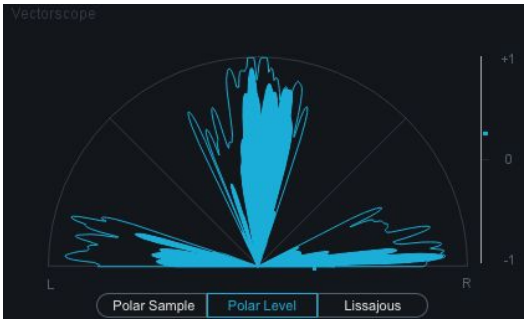


Abb. 15: Songanalyse

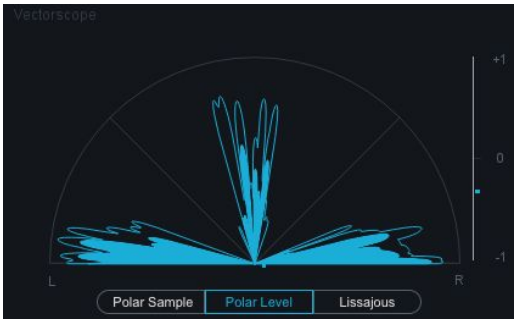


Abb.16: Seitenspeaker

Stellt man die Polarlevel-Darstellung des Vektorskops von Songfile und Seitenspeaker gegenüber, zeigt sich wieder, dass das Seitensignal über die

Seitenspeaker im Vergleich zum Mittensignal verstärkt ist.

5.5.3 Kane Brown - Be Like That (0:00 - 0:22)

Das Intro beginnt mit einer auf der E-Gitarre gespielten Akkordfolge, dazu sind stark effektierte Background-Sängerinnen zu hören. Nach 4 Takten setzen Gesang, elektronische Hihats und Claps ein. Der Gesang besteht aus Hauptstimme und 2 dezent in den Hintergrund gemischte Harmoniestimmen.
Die Gitarre als auch die effektierten Background-Stimmen sind mit Reverb versehen. Die Background-Stimmen wandern im Stereobild und sind recht stark im Seitensignal vertreten. Der Hauptgesang ist präsent im Mitten- als auch im Seitensignal und wirkt daher sehr breit.
Beim Blick auf das Correlometer lässt sich zu Beginn wieder eine Korrelation im negativen Bereich ab etwa 300 Hz feststellen, wenn auch nicht so ausgeprägt wie bei “I Hope”. Zwischen 400 Hz und 1000 Hz schlagen die Negativwerte besonders weit aus. Wenn der Gesang einsetzt, lässt sich wieder das Stereo-Korrelationsmuster von Abb. 10 erkennen.

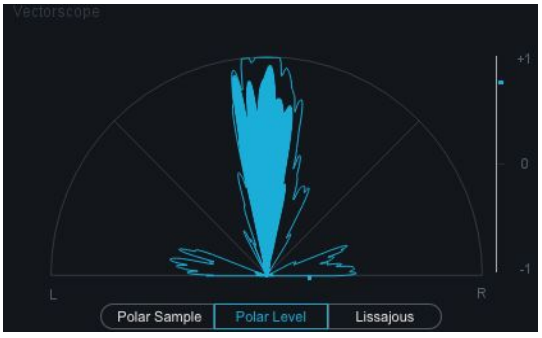


Abb. 17: Songanalyse

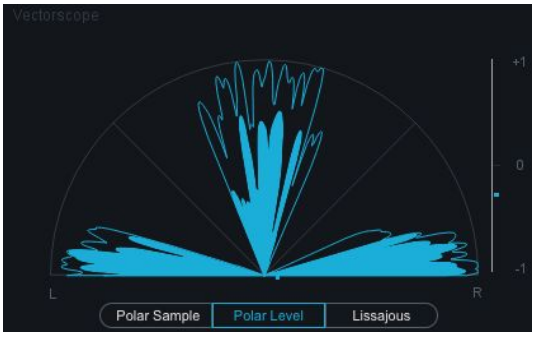


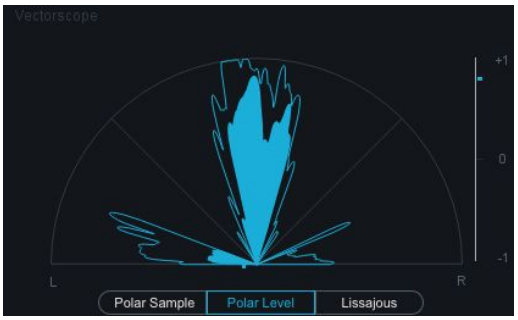
Abb. 18: Seitenspeaker

Beim Vergleich des Vektorskops fallen in diesem Beispiel größere Unterschiede auf: Das Seitensignal der Steitenspeaker ist stark ausgeprägt, während beim Songfile das Mittensignal stark dominiert.

5.5.4 Norah Jones - Don’t Know Why (0:46 - 1:11)

Die Instrumentierung des Stücks besteht aus einem Schlagzeug, das mit Jazzbesen gespielt wird, einer E-Gitarre, einem Klavier, einem Kontrabass, und der Hauptstimme, im Refrain kommen noch zwei Backgroundsängerinnen hinzu.
Die Produktion klingt organisch und kommt im Vergleich zu den bisherigen Beispielen ohne elektronische Sounds aus. Die Instrumente spielen in einem natürlich klingenden Raum und die Hauptstimme wurde mit einem dezenten Reverb versehen.

Das Klavier und die Gitarre sind recht stark im Seitensignal vertreten und demnach gut dekorreliert. Das Reverb der Hauptstimme, sowie die Backgroundstimmen sind ebenfalls im Seitensignal präsent. Im Sinne einer breiten ASW funktionieren hier vor allem die Gitarre und das Klavier gut.
Betrachtet man das Korrelationsverhalten der beiden Seitenspeaker, zeigt sich in den dekorrelierten Passagen, das sind jene bei denen die Stimme aussetzt und Gitarre und Klavier präsent sind, wieder das Korrelationsmuster des rosa Rauschens.



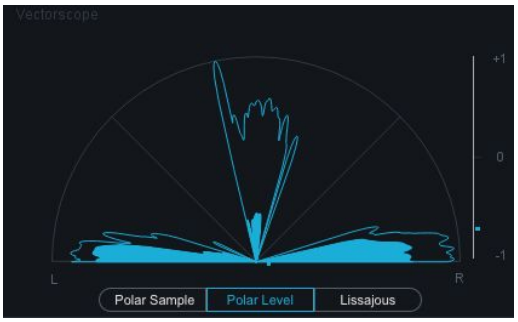


Abb.19: Songanalyse

Abb.120: Seitenspeaker

In der Vektorskop-Ansicht sieht man erneut, dass das Seitensignal bei den Seitenspeaker verstärkt wird.

5.5.5 The Weekend ft. Daft Punk - Starboy (2:09 - 2:40)

Im betrachteten Abschnitt des Songs erklingen einige Synthesizer, elektronisches Schlagzeug, Hauptgesangsstimme und eine Vocoder-Stimme.
Bezüglich einer breiten ASW stechen hier vor allem 2 Sounds heraus:
1. Die Snare, welche eine Mischung aus synthetischer Snare und Claps ist. Diese ist ab etwa 1 kHz im Seitensignal sehr präsent.
2. Ein tonal gleichbleibender Synthesizersound, der bei 400 Hz im Seitensignal hervorsticht.
Die Songpassage ist im gesamten sehr monokompatibel, über 300 Hz dekorreliert es etwas, ist jedoch im Schnitt mit einem Wert zwischen +0,5 und +0,75 noch recht korreliert. Dies erklärt womöglich dass die beiden beschriebenen Sounds hervorstechen.

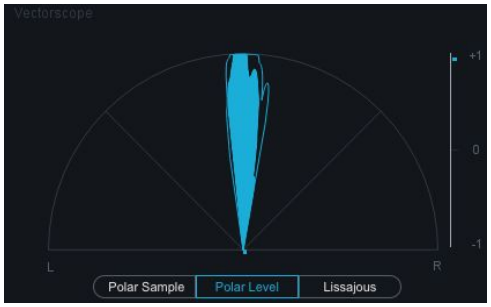


Abb. 21: Songanalyse

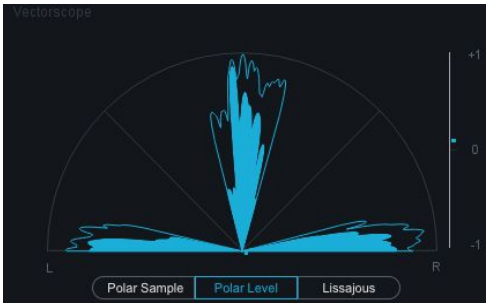


Abb. 22: Seitenspeaker

Hier ist deutlich zu erkennen, dass die analysierte Songpassage stark monokompatibel ist und dass die Seitenspeaker des Echo Studios das Signal dekorrelieren.

5.5.6 Reid Willis - Building the Monolith 3D (3:08 - 3:32)

Bei diesem Stück handelt es sich um einen binauralen Mix, der also eigentlich für die Kopfhörerwiedergabe gedacht ist, jedoch funktioniert die ausgewählte Passage auch über den Echo Studio. Im ausgewählten Abschnitt dieses experimentellen elektronischen Stücks sind synthetisch klingende “Whooshes” zu hören, die etwas nach Sci-fi klingen und langsam beschleunigen. Diese “Whooshes” werden im Raum gepanned und haben einen hohen Geräuschanteil.
Das Korrelationsmuster des binauralen Signals ist wie zu erwarten sehr wechselhaft, wobei es unter 200 Hz korreliert. Im Vergleich mit den Seitenspeaker gibt es grundsätzliche Ähnlichkeiten, allerdings gehen die Korrelationswerte hier noch weiter in den negativen Bereich.

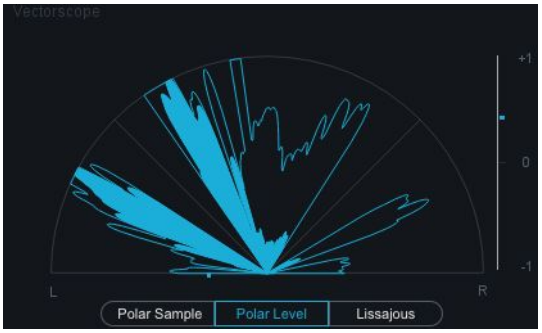


Abb. 23: Songanalyse

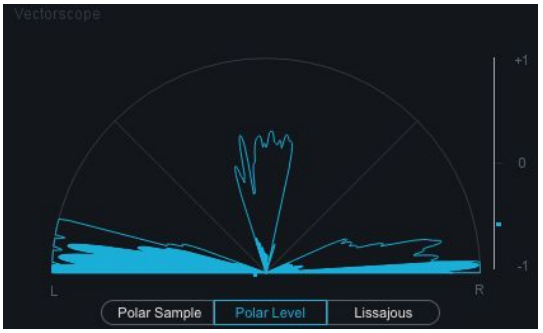


Abb. 24: Seitenspeaker

In der Vektorskopansicht zeigt sich einerseits, dass das Stereobild des Stücks im Vergleich zu den anderen Beispielen hinsichtlich der Korrelation diffuser ist, und

andererseits, sieht man wieder, dass bei den Seitenspeaker das Seitensignal deutlich präsenter ist.

5.5.7 Max Cooper - Veil of Time 3D (0:46 - 1:04)

Dieses Stück ist ebenfalls im binauralen Format und ist dem Ambient-Genre zuzuordnen. Es sind sphärische Synthesizer-Pads, langgezogene “Whooshes”, hölzernes Krachen, ein Sound der sich nach Granularsynthese anhört und eine im Hintergrund spielende Synthesizer-Melodie zu hören.
Alle Sounds sind mit viel Reverb versehen, sodass die sphärischen Sounds verschmelzen und somit eine dynamische Atmosphäre schaffen. Das hölzerne Krachen und vor allem der “Granularsynthese-Sound” tritt dabei immer wieder in den Vordergrund.
Korrelationstechnisch halten sich die Werte über 400 Hz größtenteils bei der Nuller-Marke auf. Unter 400 Hz korreliert das Signal weitgehend und ist somit insgesamt nicht weniger monokompatibel als so mancher Normal-Stereotrack.
Das Korrelationsmuster der Seitenspeaker lässt über der 400 Hz-Marke wieder das des stereo rosa Rauschens erkennen; darunter korreliert das Signal.

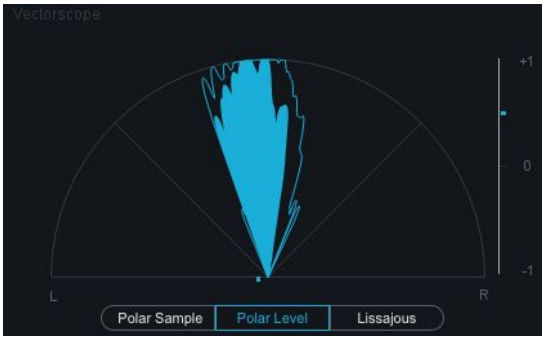


Abb. 25: Songanalyse

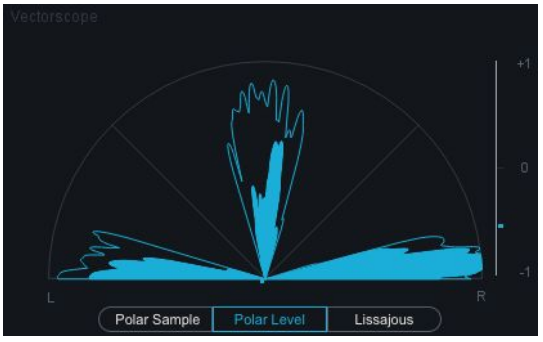


Abb. 26: Seitenspeaker

Im Vektorskop erkennt man das für ein binaurales Format sehr korreliertes Stereobild. Bei den Seitenspeaker ist abermals die Verstärkung des Seitensignals klar zu erkennen.

5.6 Abhörraum

Für das weitere Analyseverfahren ist es nötig den Abhörraum zu kennen. Dies war bei den vorhergehenden Analyseschritten nicht notwendig, da das Verhalten der Lautsprecher untersucht wurde und nicht der Raumeindruck.
Als Abhörzimmer fungierte ein Wohnraum, denn genau für solche Umgebungen ist der Echo Studio geschaffen. Der Smartspeaker ist abhängig von Reflektionen, die in einem akustisch optimierten Studioraum nur bedingt auftreten, wodurch die Technologie des Smartspeakers nur eingeschränkt zur Geltung kommen würde.

Der Raum ist fast quadratisch, mit Seitenlängen von 4,50m und 4,60m, und einer Wandhöhe von 3,10m. Die Wände sind größtenteils frei, wodurch sie als gute Reflektionsflächen für den Echo Studio dienen. Eingerichtet ist der Raum mit ein paar absorbierenden Möbelstücken wie einem Bett, zwei Sofas und einem Teppich.
Der Echo Studio wurde mittig im hinteren Drittel des Raumes platziert, zudem wurden links und rechts neben dem Smartspeaker zwei Studiomonitore platziert. Alle Speaker befinden sich mit 1m zum Boden etwa auf der Höhe der Ohren in Sitzposition. Der Abhörpunkt befindet sich mittig im vorderen Drittel des Zimmers. (siehe Abb. 27)

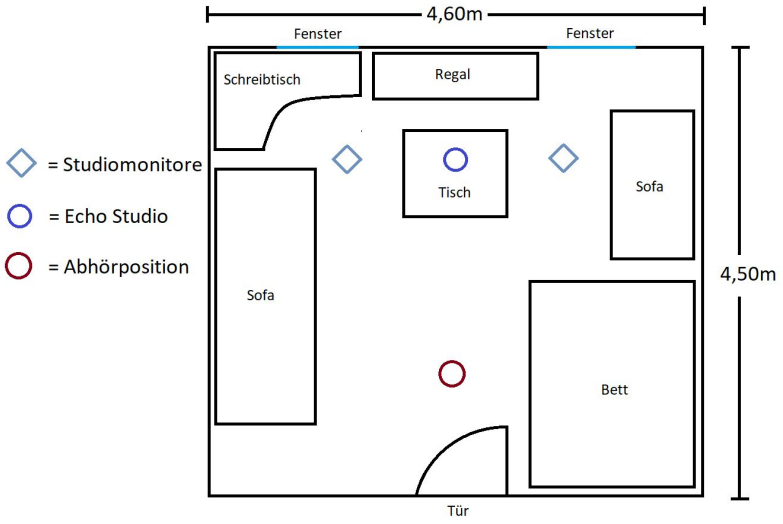


Abb. 27: Abhörraum

An der Abhörposition wurde die Nachhallzeit, als wichtige Eigenschaft für den Raumklang, nach dem RT-60-Standart gemessen. Das bedeutet, es wurde der Zeitraum gemessen den der Schall braucht um vom Impulsbeginn 60 dB leiser zu werden. (vgl. http://www.sengpielaudio.com/Rechner-RT60.htm) Hierfür wurden fünf Impulsantworten mit einem kalibrierten Messmikrofon aufgenommen und mittels der Raumakustik-Analysesoftware Room EQ Wizard (REW) ausgewertet. Üblicherweise bezieht man sich bei der Angabe der Nachhallzeit auf die Frequenzen 500 Hz oder 1000 Hz. Hier wurde zusätzlich noch eine Oktave unter und über besagten Frequenzbändern betrachtet. Es wurden somit folgende vier Frequenzen untersucht: 250 Hz, 500 Hz, 1000 Hz, 2000 Hz.
Aus den Werten der einzelnen Aufnahmen wurde der Durchschnitt für jede Frequenz berechnet, es ergeben sich Werte zwischen 0,62 und 0,89 Sekunden Nachhallzeit (siehe Tabelle, Abb. 28).

	250Hz	500Hz	1000Hz	2000Hz
Aufnahme 1	0,77s	0,84s	0,71s	0,65s
Aufnahme 2	0,72s	0,81s	0,76s	0,57s
Aufnahme 3	0,69s	1,17s	0,77s	0,58s
Aufnahme 4	0,65s	0,79s	0,68s	0,63s
Aufnahme 5	0,77s	0,84s	0,71s	0,65s
Durchschnitt	**0,72s**	**0,89s**	**0,73s**	**0,62s**

Abb. 28: Nachhallzeiten des Abhörraumes

Die gemessene Nachhallzeit ist im Normalbereich für Wohnräume und hat nachgewiesenermaßen keinen störenden akustischen Effekt, was die Musikwiedergabe über Lautsprecher betrifft.[107] Der Raum kann also als repräsentativ für die meisten Wohnräume angesehen werden und bietet somit optimale Voraussetzungen für die folgende Untersuchung.

5.7 Binauralanalyse

In den vorhergehenden Analyseabschnitten wurde die Aktivität der Speaker des Echo Studios untersucht. Im Folgenden soll nun mithilfe von Binauralaufnahmen eine Analyse der tatsächlichen Hörerfahrung erfolgen.
Die Aufnahmen wurden im oben beschriebenen Abhörraum an der in Abb. 26 eingezeichneten Abhörposition mittels der binauralen WPM-10 In-Ear-Mikrofone von Roland gemacht. Anhand dieser Aufnahmen kann durch die Untersuchung der Korrelationsmuster eine objektive Deutung auf die ASW und das LEV getätigt werden, wobei Korrelationswerte um 0 als optimal für ein immersives Hörerlebnis gelten. Allerdings sei an dieser Stelle betont, dass eine immersive Hörerfahrung von mehr Faktoren als dem Korrelationswert abhängt, wie in den einführenden Kapiteln beschrieben.
Es wurden Aufnahmen von der Wiedergabe des Echo Studios mit aktivierter, sowie deaktivierter Upmixfunktion, als auch von Studiomonitoren untersucht. Die Studiomonitore wurden herangezogen um eine größere Vergleichsbasis zu haben und um zu sehen, was die Unterschiede zu dieser ‘normalen’ Stereowiedergabe sind.
Subjektiv empfunden ist der auditive Einhüllungsgrad beim Echo Studio mit aktivierter Raumklangfunktion am höchsten, Ziel der Analyse ist es diese Empfindung auf eine objektive Ebene zu bringen und diese gegebenenfalls zu validieren.

[107] vgl. Rampelmann (2015), S.9

Zunächst wurde rosa Rauschen in mono und in stereo analysiert, um ein Korrelationsmuster eines spektral kontinuierlichen Signals zu generieren. Denn so lassen sich Unterschiede am besten feststellen. Zusätzlich wurden für den praktischen Bezug wieder die gleichen Musikbeispiele wie im vorigen Abschnitt betrachtet.

Die Analyse des rosa Rauschens in stereo zeigt niedrigere Korrelationswerte bei aktivierter Raumklangfunktion, speziell im Bereich von etwa 400 Hz - 3 kHz , mit Negativ-Peaks von etwas über -0,5 bei 500 Hz - 1 kHz. Bis 500 Hz herrscht bei beiden Wiedergabefunktionen des Echo Studios eine starke Korrelation, was damit zu erklären ist, dass die tiefen Frequenzen nur von einem Speaker wiedergegeben werden. Bei der Wiedergabe über die Studiomonitore zeigt sich ein etwas anderes Muster: Die unteren Mitten und Bässe sind nicht so korreliert wie beim Echo Studio, zeigen aber bis 700 Hz deutlich positive Korrelationswerte. Die dekorrelierteren tiefen Frequenzen ergeben sich aufgrund der 2 Schallquellen, die beim Echo Studio, wie erwähnt, nicht vorhanden sind. Bei 2 kHz gibt es einen Negativ-Peak von -0,25. Dies unterscheidet sich vom Echo Studio, wo der Negativbereich tiefer liegt und vor allem bei aktiver Raumklangfunktion stärker ausgeprägt ist. (Abb. 29)

Abb. 29: Rosa Rauschen stereo. Links: Echo Studio Upmix-Funktion, Mitte: Echo Studio normal, Rechts: Studiomonitore

Bei der Wiedergabe des rosa Rauschens in mono über den Echo Studio zeigt sich ein positiveres Korrelationsmuster als beim Stereo-Rauschen, sowohl bei aktiver, als auch bei inaktiver Raumklangfunktion. Gegenüber der inaktiven Upmix-Funktion, weist die Wiedergabe mit Raumklangfunktion im Bereich 2,5 kHz - 8 kHz eine deutlich höhere Korrelation auf, mit einem Peak bei 4 kHz. Dies könnte vor allem auf den Frontalspeaker zurückzuführen sein, der bei deaktivierter Upmix-Funktion inaktiv ist. Bei 10 kHz gibt es jedoch eine auffällige Dekorrelation, die möglicherweise durch Interferenzen vom direkten Signal des Frontalspeakers und den Reflektionen der Seitenspeaker zustande kommt.
Die Studiomonitore weisen im Vergleich zur Wiedergabe des rosa Rauschens in

stereo kein positiveres Korrelationsmuster auf, stattdessen gibt es zwei Negativ-Peaks bei 1,5 kHz und 4 kHz.

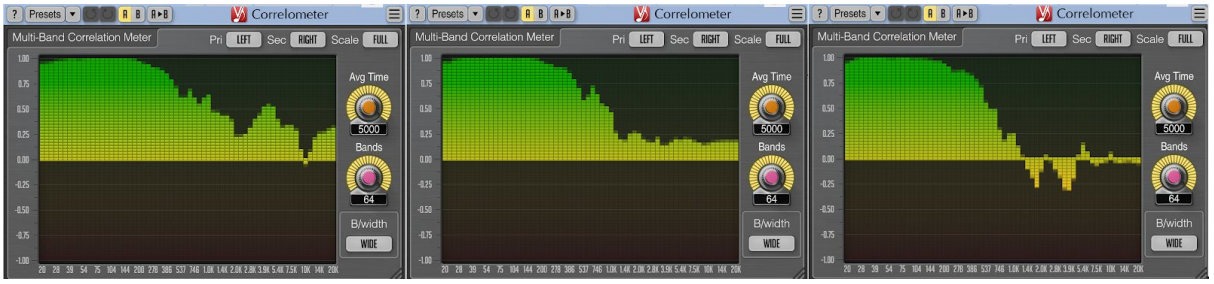


Abb. 30: Rosa Rauschen mono. Links: Echo Studio Upmix-Funktion, Mitte: Echo Studio normal, Rechts: Studiomonitore

Die Korrelationsmuster der Musikbeispiele sind aufgrund ihrer “Lebendigkeit” wie zu erwarten geprägt von Fluktuationen. Trotzdem können die einzelnen Wiedergabemethoden in ihrer Korrelation von einander unterschieden werden. So zeigen sich die beim stereo rosa Rauschen beobachteten Muster auch in den Musikbeispielen. Wie klar diese zu erkennen sind hängt vom Korrelationsgrad der Produktion ab: Am Beispiel von Trentemøller ist wie beim rosa Rauschen bei aktivierter Upmix-Funktion ein negativer Korrelationsbereich um 1 kHz ersichtlich, wenn auch nicht ganz so ausgeprägt. Ebenso gibt es eine starke Korrelation bis 500 Hz. Die Studiomonitore zeigen ebenfalls wie beim rosa Rauschen einen Negativ-Peak um 2 kHz.
Das Beispiel von Norah Jones hingegen ist von Seiten der Produktion im Gesamten wesentlich korrelierter als das Trentemøller-Beispiel. Dies ist ist auch im Korrelationsmuster der Binauralaufnahme ersichtlich: Bei aktivierter Upmix-Funktion gibt es wie beim mono rosa Rauschen einen positiven Korrelations-Peak. Bei den Passagen in denen das Klavier, welches im Seitensignal recht präsent ist, Phrasen einwirft, kommt das vom stereo Rauschen bekannte Korrelationsmuster wieder zum Vorschein.

Die Binauralanalyse hat gezeigt, dass die Raumklangfunktion Stereosignale am stärksten dekorreliert. Dazu weisen wiedergegebene korrelierte Signale bei aktiver Raumklangfunktion die höchste Korrelation auf, das heißt also, dass die Differenz von korrelierten und dekorrelierten Signalen, die über den Echo Studio mit aktivierter Upmix-Funktion wiedergegeben werden, am größten ist. Das bedeutet für die Hörerfahrung wiederum, dass die Klangbreite, also die ASW, dynamischer agieren kann, was für ein immersives Erlebnis sicher nicht von Nachteil ist.

5.8 Diskussion Analyse

Die Wiedergabeanalyse von Monosignalen und Stereosignalen haben gezeigt, dass je nach Signaltyp unterschiedliche Lautsprecher aktiv sind, bzw. die Intensität der Speaker variiert. Auffällig ist vor allem das Aktivitätsmuster des Frontalspeakers und des Vertikalspeakers: Bei (breitbandigen) Monosignalen ist der Frontalspeaker sehr prominent und der Vertikalspeaker kaum aktiv, bei (breitbandigen) Stereosignalen ist es umgekehrt. Um diesen Zusammenhang weiter zu bestätigen wurde eine kontinuierliche Verminderung der Stereobreite des Rauschens bis hin zu Mono vorgenommen. Die Hüllkurven der Aufnahmen des Vertikal- und Frontalspeakers zeigten die stetige Aktivitäts-Zunahme bzw. -Abnahme wie vermutet.
Man kann nun also schließen, dass die Intensität von Frontal- und Vertikalspeaker abhängig vom Korrelationsgrad des wiedergegebenen Signals ist, wobei bei einem Korrelationsgrad von 1 (Mono) der Frontalspeaker dominiert und bei 0 (maximale Dekorrelation) der Vertikalspeaker.
Zusätzlich hängt die Aktivität des Frontalspeakers vom Panning ab. Bei aktiviertem Raumklang funktioniert das Panning durch eine Mischung aus Phasenverschiebung und Spektraländerung.

Die Songanalyse hat gezeigt, dass die Upmixfunktion des Echo Studios von einem dekorrelierten Mix profitiert. Zudem ist der Einsatz von Reverb für das LEV essentiell. Die Vektorskopanalyse hat bei allen Beispielen gezeigt, dass die Seitenspeaker das Signal weiter dekorrelieren. Man darf an dieser Stelle jedoch nicht vergessen, dass der Frontalspeaker das Mittensignal vor allem in den Höhen stützt, was die Gesamt-Abstrahlung des Echo Studios korrelationstechnisch wieder etwas ausgeglichener erscheinen lässt.
Weiters war zu beobachten, dass bei den gut dekorrelierten Songpassagen das Korrelationsmuster, das sich bei der Analyse des rosa Rauschens gezeigt hat, (siehe Abb. 9) ebenfalls zu erkennen ist. Dieses Muster zeichnet sich durch eine starke negative Korrelation im Bereich von 350 Hz bis 7,5 kHz aus. Laut Pfanzagl-Cardone sind die Frequenzbereiche um 500 Hz, 1000 Hz und 2000 Hz hinsichtlich der ASW besonders relevant.[108] Diese Bereiche sind auch beim Korrelationsmuster des rosa Rauschens am stärksten ausgeprägt.
Die Analyse der beiden binauralen Songbeispiele (Kapitel 5.5.6 und 5.5.7) hat gezeigt, dass dieses Format durchaus einen positiven Effekt hinsichtlich des LEVs bei der Wiedergabe über den Echo Studio hat. Dies gilt vor allem für atmosphärische Klänge. Da diese Sounds ohnehin diffus sind, wirkt sich der Übersprechungseffekt nicht negativ auf das Klangbild aus, trotzdem können so gewisse Richtungs- und Raumeindrücke entstehen.

[108] vgl. Pfanzagl-Cardone (2010), S.32

Bei der Binauralanalyse wurde beobachtet, dass die Differenz der Korrelationswerte zwischen Mono- und Stereosignalen bei aktivierter Upmix-Funktion am größten ist. Die Studiomonitore haben dagegen sowohl bei Stereo- als auch bei Monosignalen ein dekorreliertes Muster gezeigt und konnten die Raumreflektionen also nicht zu ihren Vorteil nutzen. Das heißt, dass der Echo Studio in akustisch unbehandelten Wohnräumen zusätzlich zur besseren Einhüllung ein definiertes Stereobild, was die Differenzierung von Mitten- und Seitensignal betrifft, bietet.

Zusammenfassend hat die Analyse des Echo Studios gezeigt, dass dieser mit aktivierter Raumklangfunktion bei folgenden Audio-Produktionseigenschaften hinsichtlich einer immersiven Hörerfahrung gut funktioniert:

- Bei dekorrelierten Produktionen und Sounds
- Bei spektral dichten atmosphärischen Klängen
- Bei Reverb und Delay
- Bei binauralen Atmosphären/Pads

Die gewonnenen Analyseerkenntnisse dienen als Grundlage für den folgenden praktischen Teil, der Produktion von Audiocontent für den Echo Studio.

6. Produktion

In folgendem Kapitel sollen anhand der Dokumentation des Produktions-Workflows eines für den Echo Studio produzierten Stücks weitere Erkenntnisse über dessen Upmix-Funktion und deren Nutzung aus produktionstechnischer Sicht gewonnen werden.
Dank der vorhergehenden Analysearbeit konnten bereits substantielle Richtlinien für die Produktion aufgestellt werden:

Dekorrelation: Dass das Konzept von Korrelation und Dekorrelation maßgeblich für räumliche Hörerfahrungen verantwortlich ist, wurde nicht nur theoretisch erörtert, sondern konnte anhand der Analyse des Echo Studios auch empirisch nachgewiesen werden. Für die Audioproduktion gibt es 3 unterschiedliche Konzepte die Stereobreite zu vergrößern und demnach den Korrelationsgrad zu steigern:
1. Delay Processing: Der Stereotrack wird matriziert, worauf kurze unterschiedlich lange Delayzeiten genutzt werden um Erste Raumreflektionen nachzuahmen.
2. Komplementäre Filterung: Nach der Matrizierung des Signals wird dieses im Frequenzbereich durch komplementäre Filter gesplittet und später mit einem anderen stereophonen Verhältnis wiedervereint.

3. M-S Processing: Das Verhältnis von Mittensignal und Seitensignal wird geändert, sodass das Seitensignal hervorgehoben wird.[109]

Atmosphäre und spektrale Dichte: Harmonisch dichte und diffuse Klänge tragen zur Verbesserung des LEVs bei, welches durch die späteren diffusen Reflektionen eines Schallereignisses definiert ist.[110] Um ein wahrgenommenes diffuses Schallfeld zu erzeugen, sollte im Bereich von 125 Hz bis 8 KHz ein niedriger Korrelationsgrad zwischen den Kanälen bestehen.[111] So erhält die Audioproduktion durch transientenarme atmosphärische Sounds einen einhüllenden Charakter. Die Wiedergabe über den Echo Studio mit aktivierter Raumklangfunktion profitiert bei solchen Sounds vor allem durch den Vertikalspeaker.

Reverb und Delay: Durch diese Effekte profitieren vor allem transientenreiche Sounds hinsichtlich einer Einhüllung bzw. der Klangbreite. Wie bei atmosphärischen Sounds trägt der diffuse Signalanteil, der durch die Verwendung eines Reverbs entsteht, zu einer immersiveren Hörerfahrung bei. Durch Delay-Effekte lassen sich Raumeigenschaften wie erste Reflektionen oder Echos simulieren.

Binaurale Atmosphären/Pads: Obwohl binaurale Signale für die Wiedergabe über Kopfhörer konzipiert sind, konnten über den Echo Studio positive Effekte hinsichtlich einer räumlichen Hörerfahrung beobachtet werden. Flächige Sounds (Pads) zeichnet ein diffuser Klangcharakter aus, was wiederum bedeutet, dass nicht die Richtwirkung relevant ist, sondern die Verteilung im Raum. Das bedeutet das Problem des Übersprechungseffekts bei der Wiedergabe von binauralen Signalen ohne Kopfhörer ist in diesem Fall vernachlässigbar.

Neben den genannten Richtlinien sollen für die Audioproduktion weitere Konzepte und akustische Eigenschaften bedacht werden:

Bewegung und Statik: Gerriet Sharma hat sich mit der Komposition von Klangmaterial für den Ikosaeder-Lautsprecher vom Grazer Institut für elektronische Musik beschäftigt, der mittels Beamforming dreidimensionale Hörerfahrungen schaffen kann. Sharma teilt Raum-Klangphänomene in 3 Ebenen ein: Ebene 1 sind statische im Raum wahrgenommene Klänge. Ebene 2 sind zeitvariante Klänge mit bestimmten Bewegungsmustern. Ebene 3 stellt die Überlagerungen verschiedener Phänomene der beiden anderen Ebenen dar.[112]
Auch wenn bei der Stereo-Wiedergabe über den Echo Studio nur begrenzt die Möglichkeit besteht Klangobjekte im Raum zu platzieren, sollte bei der Produktion, im Sinne der 3 genannten Ebenen, auf Bewegung und Statik von Klängen geachtet

[109] vgl. Baskind, u.a. (2015), S.5
[110] vgl. Lee (2013), S.1
[111] vgl. Cousins, Bleeck Fazi (2017), S. 8
[112] vgl. Sharma (2016), S.106/107

werden.

Psychoakustik: Bei der Ortung von Klängen spielt auch die Vertrautheit eine Rolle, so sind gewohnte Sounds leichter zu lokalisieren, da wir ein gewisses Vorwissen über deren Spektrum haben und somit wissen inwiefern der Klang gefiltert wurde.[113] Durch die Verwendung von vertrauten Klängen kann unser Gehör also bis zu einem gewissen Grad ausgetrickst werden, was nützlich für die begrenzten räumlichen Möglichkeiten einer Stereowiedergabe über Lautsprecher sein kann.

Lautsprecher: Für die Produktion sollte außerdem im Hinterkopf behalten werden, dass die Abstrahlcharakteristik von Lautsprechern im Normalfall stark frequenzabhängig ist. Hohe Frequenzen werden dabei gerichtet abgestrahlt, und tiefe Frequenzen annähernd omnidirektional.[114] Das heißt es wird wenig Sinn machen den Bass in irgendeine Richtung zu pannen. Dazu kommt im Fall des Echo Studios, dass dieser sowieso nur einen Tieftöner besitzt.

6.1 Die Komposition “Room Switch”

Folgendes Kapitel setzt sich mit der Produktion des Stückes “Room Switch” auseinander, welches für die Wiedergabe über den Echo Studio komponiert wurde. Das Stück lässt sich in 6 distinkte Formteile gliedern, wodurch ein großes Spektrum an unterschiedlichen Klangkomponenten zum Einsatz kommt. Dadurch ergeben sich für den/die Hörer*in mehr Eindrücke um die Wirkungsweise des Smartspeakers zu beurteilen.
Produktionstechnisch wurden die im vorigen Kapitel näher beschriebenen Punkte, welche sich aus der Analysearbeit und Recherche etablierten, praktisch umgesetzt. Es wurde also darauf geachtet unterschiedliche Möglichkeiten der Dekorrelation zu nutzen, atmosphärische Sounds zu nutzen, mit Reverb und Delay zu arbeiten, binaurale Komponenten zu nutzen, auf Bewegung zu achten, als auch an psychoakustische Effekte zu denken.
Konzeptionell aus der Sicht des Storytellings können die einzelnen Formteile als unterschiedliche Räume wie beispielsweise in einem Museum betrachtet werden. Denn in Teil A hört man einen Protagonist durch einen Raum gehen, welcher nach und nach Schalter betätigt. Jene Schalter markieren in der Komposition manchmal den Wechsel in den nächsten Formteil, oder bildlich gesprochen in den nächsten Raum. Diese Metapher passt auch insofern gut, als dass sich die Formteile recht stark voneinander unterscheiden, wie eben Räume in einem Museum. Trotzdem wurde darauf geachtet, dass das Stück eine gewisse Kohärenz behält, indem gewisse Themen oder Klangkomponenten in anderen Formteilen wieder aufgegriffen

[113] vgl. Wenzel, Begault, Godfroy-Cooper (2017), S.19
[114] vgl. Pfanzagl-Cardone (2010), S.65

werden.
Das Stück, sowie eine Kurzfassung davon, als auch die einzelnen Audiospuren können unter dem auf Seite 68 angeführten Link angehört und heruntergeladen werden.

6.1.1 Teil A (0:00 - 0:13)

Das Stück beginnt mit dem öffnen einer Tür, anschließend hört man Schritte. Kurz darauf bleibt die Person stehen und betätigt einen Schalter und plötzlich ist ein Schlagzeug-Groove zu hören. Nach kurzer Zeit wird der Schalter erneut betätigt und das Schlagzeug verstummt abrupt. Die Person meldet sich mit roboter-artiger Stimme zu Wort und sagt: “Mono is cool, but I want stereo!” Die Person geht zwei Schritte weiter und betätigt den nächsten Schalter, worauf sich der gehörte Raum öffnet und zu Stereo wird.
Teil A ist zur Gänze in mono gehalten und soll einen starken Kontrast bezüglich der empfundenen Räumlichkeit zum nächsten Formteil etablieren. Es wurden möglichst trockene Aufnahmen verwendet um diesen Unterschied zu verdeutlichen.

6.1.2 Teil B (0:13 - 1:20)

Dieser Teil beginnt mit der Betätigung des dritten Schalters, worauf sich die Komposition räumlich öffnet, indem von Mono auf Stereo gewechselt wird. Zudem wird das Stück ab hier musikalisch.
Teil B zeichnet sich durch einen langsamen und atmosphärischen Charakter aus. Um auf die Produktion näher einzugehen, sollen die einzelnen Klangkomponenten näher betrachtet werden:

Synthesizer Pads: Das atmosphärische Fundament bilden die beiden Pad-Sounds, welche mit dem Synthesizer-Plugin “Massive” von Native Instruments synthetisiert wurden. Die beiden Sounds zeichnet ein dichtes Frequenzspektrum aus, wodurch sich über den Echo Studio eine gute Einhüllung ergibt. Es wurden beide Pads mit einem Reverb versehen um die Diffusität der Sounds zusätzlich zu steigern. Weiters wurde ein LFO gesteuerter Highcut-Filter für jeweils ein Pad genutzt, dadurch wirken die Sounds etwas ‘lebendiger’. Mit dem binauralen DearVR pro Panning-Plugin wurden beide Pads in unterschiedliche Richtungen gepanned. Die Sounds wurden so ausgerichtet, dass die räumliche Wirkung subjektiv empfunden am besten ist, was bei etwa 40° vorne links bzw. rechts der Fall war. Beide Sounds wurden zu einer Gruppe zusammengefasst, auf diese wurde für etwas mehr Bewegung ein automatisierter Highcut-Filter gelegt.
Die beiden Pads haben im Schnitt einen Korrelationswert von -0,25, mit einem Negativ-Peak bei 700 Hz.

Bass: Der Basssound ist ebenfalls synthetisch erzeugt und ist in mono. Wie schon erwähnt, macht es aus zwei Gründen keinen Sinn Bässe zu dekorrelieren: 1. Tiefe Frequenzen breiten sich über Lautsprecher ohnehin omnidirektional aus und 2. besitzt der Echo Studio nur einen Subwoofer.

Schlagzeug: Da es sich bei folgenden Klängen auch um synthetische handelt, stimmt die Bezeichnung “Schlagzeug” eigentlich nicht. Der Einfachheit halber wird der Begriff trotzdem für die perkussiven Sounds genutzt.
Die Snare-Drum wurde mittels Delay Processing dekorreliert, das heißt es wurde das Sample dupliziert und um ein paar Millisekunden phasenverschoben. Der entstandene Sound wurde mit einem Ping-Pong-Delay effektiert. Da das Delaysignal eine stärkere Externalisation aufwies, wurde dieses anschließend extra aufgenommen. Die Aufnahmen des Delaysignals wurden erneut phasenverschoben. Aufgrund des Haas-Effekts wird das entstehende Audio-Sample von jener Seite stärker wahrgenommen, wo die Phase zuerst beginnt. Es wurden also zwei Snare-Samples gemacht, eines wird stärker von rechts wahrgenommen und eines von links. Kompositorisch wechseln sich beide Sounds ab. Der dezent nach rechts hängende Snare-Sound wurde mit einem Reverb versehen und das andere Snare-Sample blieb hingegen trocken.
Bei der Hihat wurde wie bei der Snare vorgegangen, wieder gibt es ein Sample das links und eines das rechts wahrgenommen wird, hier jedoch deutlich stärker.
Die Kickdrum ist nicht dekorreliert worden, jedoch mit einem Delay effektiert.

Synth Lead: Dieser Sound setzt im letzten Drittel von Teil B ein. Zur Erzeugung wurde wieder der Massive-Synthesizer verwendet. Frequenztechnisch ist der Klang mittenlastig, sodass er sich im Mix gut durchsetzt. Nach vier Takten setzt eine Harmoniestimme ein und die Melodie wird variiert, zudem wird der Sound mit einem LFO-automatisierten High-Cut-Filter moduliert. Die ASW wurde mit einem Chorus-Effekt und einem Flanger vergrößert. Mit den Effekten hat der Sound einen Korrelationswert von 0,25, jedoch mit Fluktuationen aufgrund des Flangereffekts.

Rosa Rauschen: Mit dem Schlagzeug setzt ein rosa Rauschen ein, welches mit einem Lowcut bei 250 Hz und einem Flanger-Effekt versehen wurde. Durch den Einsatz des Flangers, der recht extrem eingestellt wurde, ergeben sich oszillierende Tonhöhenänderungen die sich im Raum gut verteilen. Die Höhenanteile des Rauschens werden über den Vertikal-Speaker gut auf die Decke gestrahlt, sodass diese von oben wahrgenommen werden.

Reversed Glass: Eine weitere Klangkomponente besteht aus rückwärts abgespielten und teils stark tonhöhenvariierten Klängen von zerbrechendem Glas. Da diese Sounds spektral dicht, also geräuschhaft sind, wurde wieder mit dem DearVR pro-Panner versucht Richtungseindrücke zu erzeugen. Am besten

funktionierte das mit der Plugin-Funktion “reflektions only” bei der nur die Raumreflektionen bewegt werden. Dies resultiert gleichzeitig in einem diffuseren Klangbild, wodurch die Sounds besser vom Speaker externalisiert wahrgenommen werden.
Später im selben Formteil werden kurze Ausschnitte der Glassounds als Hihat-Instanz verwendet. Da die Glassamples gut externalisiert wurden, war die Überlegung, dass auch daraus geschnittene kurze Samples ähnlich wahrgenommen würden. Wie sich herausstellte, funktionierte das jedoch nicht so wie erhofft. Dies lässt sich wohl dadurch erklären, dass die kurzen Samples Transienten aufweisen, welche generell näher am Speaker wahrgenommen werden.

Synth Hintergrund: Gleichzeitig mit dem Schlagzeug setzt eine Synthesizer-Melodie ein die mit viel Reverb leise im Hintergrund zu hören ist. Musikalisch wird hier eine Melodie angedeutet die erst später wieder in der Komposition auftaucht. So soll trotz der markanten Unterschiede der Formteile eine Kohärenz entstehen.

Oszillierender Sound: Am Beginn des Formteils, setzt gleichzeitig mit dem Pistolenschuss ein oszillierender Klang ein, der mithilfe des Sample-basierten Synthesizer-Plugin Iris 2 von iZotope und weiteren Effekten kreiert wurde. Dieser wurde anschließend mit dem binauralen Panner in eine Richtung bewegt.

Aus produktionstechnischer Sicht stellt Teil B wohl den experimentellsten Formteil der Komposition dar, da hier versucht wurde möglichst jede Klangkomponente zu externalisieren. Dies schlägt sich in einem durchschnittlichen Korrelationswert von -0,25 nieder, wodurch es bei einer Monowiedergabe, welche auch heutzutage für Musikproduktionen noch zu bedenken ist, zu Auslöschungen kommt. Für eine kommerzielle Produktion wäre dies zu vermeiden.
Zu bedenken ist jedenfalls auch, ob es überhaupt notwendig ist beispielsweise das Schlagzeug zu externalisieren. Zwar kann dies durchaus ein gewünschter Effekt sein, jedoch sollten gewisse Konventionen möglicherweise gewahrt werden; z.B. Kick, Snare und Bass in der Mitte.

6.1.3 Teil C - Übergang (1:20 - 1:36)

Teil B mündet abrupt in den nächsten Formteil, welcher wieder mit einem Schalter-Geräusch eingeleitet wird. Mit der Betätigung dieses Schalters kollabiert das breite Stereobild wieder zu mono. Wie schon bei Teil A soll der/die Hörer*in durch diesen starken Kontrast der Klangbreite für den räumlichen Klang des Echo Studios sensibilisiert werden. Es wurde versucht diesen Kontrast kompositorisch noch zu verdeutlichen: Nicht nur ändert sich die Stereobreite abrupt, sondern auch die Klangkomponenten: Viele breite teils atmosphärische Sounds wechseln zu

einem trockenen klar zuzuordnenden Schlagzeugrhythmus. So wirkt der Wechsel auch aus psychologischer Sicht stark, da man sich als Hörer plötzlich nur mehr auf eine Klangkomponente konzentrieren muss.
Nach dem Schlagzeugrhythmus setzt eine E-Gitarre ein und ein Build-Up in den nächsten Teil erklingt.

6.1.4 Teil C (1:36 - 2:08)

Mit dem Einstieg in den nächsten Formteil wechselt das Klangbild von mono wieder auf stereo. Während Teil B, wie erwähnt, aus produktionstechnischer Sicht experimenteller war, ist Teil C konventioneller. Eine genauere Betrachtung der Klangkomponenten soll einen besseren Eindruck über die Produktion vermitteln:

Schlagzeug: Hier wurde im Vergleich zum vorhergehenden Formteil mit echten Schlagzeugsamples gearbeitet. Die Snare-Drum wurde mit einem kurzen stereo Slap-Delay etwas verbreitert. Auf die Summe wurde etwas Raum mit einem Reverb hinzugefügt, ansonsten ist das Schlagzeug konventionell mit mittiger Snare- und Kick-Drum gehalten. Es sei noch erwähnt, dass auf den Overhead-Samples grundsätzlich schon Raumanteile vorhanden waren. Generell wirken Becken aufgrund ihrer spektralen Eigenschaften von den Schlagzeugsounds meistens am besten im Hinblick auf die räumliche Wiedergabe mit dem Echo Studio.

E-Gitarre: Die Gitarre wurde doppelt aufgenommen und jeweils 50% nach links und rechts gepanned. Mit einem Widening-Plugin wurde die Stereosumme dezent dekorreliert. Durch diese Stereoverbreiterung setzt sich die Gitarre etwas besser im Mix durch, jedoch sollte man es mit solchen Tools nicht übertreiben, da es leicht zu einer unerwünschten Änderung der Klangfarbe kommen kann. Zusätzlich wurde die Gitarre durch den gleichen Reverb gesendet wie das Schlagzeug um eine räumliche Kohärenz zu erzeugen.

Bass 2: Der Basssound wurde synthetisch erzeugt und ist konventionell in mono gehalten. Im Laufe des Formteils wird der Sound mithilfe eines Bit-Crushers und eines Saturation-Plugins moduliert.

Synth Lead: Siehe Teil B. Der Sound wurde durch denselben Reverb wie die Gitarre und das Schlagzeug gesendet.

Synth Lead 2: Der mit dem Tyrell 6 Synthesizer von u-he erzeugte Sound setzt nach 4 Takten ein. Anfangs wird die Melodie unisono im Oktavabstand von zwei Synthesizer-Instanzen gespielt, nach der ersten Wiederholung setzt eine dritte Stimme ein und spielt Terz-Harmonien dazu. Alle drei Stimmen sind leicht unterschiedlich gepanned um die ASW zu vergrößern. Zusätzlich wurde jede Spur mit dem Tube Modulator Plugin von Audiority effektiert. Dieses Vibrato- und

Autopan-Plugin wurde benutzt um dem Sound einerseits mehr Ausdruck zu verleihen und andererseits um etwas Bewegung zu erzeugen.

Rosa Rauschen: Siehe Teil B. Das rosa Rauschen setzt gleichzeitig mit dem Synth Lead 2 ein und bringt mehr Bewegung in den Mix.

Am Ende des Formteils werden alle Sounds wieder zu mono gebündelt, dieses Mal jedoch nicht abrupt, sondern mittels Fade. Gleichzeitig mit dieser Bündelung erklingt ein “Whoosh”-Sound der in den nächsten Teil mündet.
Im Gesamten betrachtet weist Teil C ab 250 Hz Korrelationswerte zwischen 0,25 und 0,5 auf, wobei ab der zweiten Hälfte des Formteils im Bereich zwischen 500 Hz und 2000 Hz Durchschnittswerte um 0 zu beobachten sind.
In Teil C wurde auf atmosphärische Klänge verzichtet und auch insgesamt ist die Produktion eher trocken und direkt gehalten. Trotzdem ist der Unterschied hinsichtlich der Räumlichkeit zum vorhergehenden Übergang in mono groß.

6.1.5 Teil D (2:08 - 2:50)

Dieser Formteil bildet wieder einen Kontrast zum vorigen Teil, indem hier überwiegend mit atmosphärischen Klängen gearbeitet wurde. Zudem gibt es hier keine Rhythmus-Instanz. Generell weist Formteil D einen Soundscape-Charakter auf der sich in der zweiten Hälfte in eine Filmmusik-Ästhetik wandelt.
Der Abschnitt beginnt in mono mit einem Geräusch des abtauchens in Wasser. Zunehmend öffnet sich das Klangbild wieder und in der Mitte des Teils bei 2:29 min. taucht ‘man’ wieder auf. Daraufhin setzen Bläser ein die eine Variation der Melodie des Synth Lead 2 in Formteil C spielen.

Zu den Klangkomponenten gehören:

Bass 2: Wie in Teil C, jedoch durch High-Cut bei 230 Hz weniger Mittenanteile.

Under Water: Eine Unterwasseraufnahme bildet das Fundament des Formteils. Aufgrund der spektralen Beschaffenheit der Aufnahme kommt es leicht zu Maskierungen mit dem Bass. Entsprechendes EQing wirkt dem entgegen.

Binaural: Nach dem Beginn des Teils entfaltet sich langsam eine binaurale Aufnahme von einem Maisfeld. Zu hören sind trockene Blätter die sich im Wind bewegen. Die vor allem in den Höhen präsenten Klänge bilden einen Kontrast zu den Unterwassersounds, sowohl frequenztechnisch, als auch räumlich.

Synthesizer Pads: Wie in Teil B. Ab der zweiten Hälfte des Formteils, gleichzeitig mit dem Auftauchen, werden die Pad-Sounds recht stark low-gecuttet, sodass die erklingenden Bläser und Streicher besser durchkommen. Grundsätzlich wurde bei

den Pads, wie schon in Teil B, darauf geachtet, dass durch automatisiertes Filtering und Lautstärke eine gewisse Lebendigkeit entsteht.

Bläser und Streicher: Die Bläser, bestehend aus Hörnern und Tuben, als auch die Streicher, bestehend aus Violinen und Violen, wurden im Seitensignal mittels High-Shelf-EQ etwas betont, sodass diese etwas breiter wirken.

Kick-Drum: Mit den Bläsern und Streichern setzt die Kick-Drum von Teil B ein.

Der in mono gehaltene Anfang von Teil D hat gezeigt, dass räumliche Kontraste nicht unbedingt nur von Mono und Stereo abhängig sind. Der Umbruch von Teil C auf D wirkt aus räumlicher Sicht nicht so spektakulär wie der Übergang von Teil B auf Teil C. Dies lässt sich durch die basslastige Qualität des Beginns von Teil D erklären, da sich tiefe Frequenzen annähernd omnidirektional ausbreiten. Deshalb wirkt der Anfang von Teil D wesentlich einhüllender als der Übergang zu Teil C, obwohl beide mono sind. Man sieht also, dass die Komposition und das Arrangement hinsichtlich der Erzeugung von räumlichen Kontrasten entscheidend ist.

6.1.6 Teil E (2:50 - 3:33)

Wie Teil D öffnet sich auch Formteil E sukzessive von mono zu stereo, dies betrifft zumindest alle musikalischen Komponenten. Denn direkt beim Übergang erklingt eine Helikopteraufnahme in stereo. Die Überlegung dahinter war es der Stereoaufnahme durch die darunterliegenden Mono-Komponenten mehr Platz zu geben. Die Helikopter-Aufnahme wurde gewählt um einen psychoakustischen Aspekt in der Komposition zu haben. Wie schon erwähnt, gibt es bei der menschlichen Klanglokalisation auch eine psychoakustische Komponente, die es uns ermöglicht bekannte Sounds leichter zu orten. Da der Klang eines vorbeifliegenden Helikopters wohl bei jedem als Höheneindruck gespeichert ist, sollte dieser auch wiedergegeben über den Echo Studio dementsprechend von oben wahrgenommen werden.
Teil E hat grundsätzlich einen Build-up-Charakter der immer atmosphärischer wird und schließlich in den den letzten Formteil mündet. Zu den Sounds gehören:

Synthesizer Pads 2: Die Pad-Sounds wurden mit dem Serum-Synthesizer von Xfer erzeugt und bestehen aus drei Layer. Es wurden obertonreiche Sägezahnwellen verwendet um den Atmosphären eine gewisse Dichte zu verleihen. Im Laufe des Formteils nehmen die Sounds an Lautstärke und Stereobreite zu. Die Pads haben einen Korrelationswert von 0.

Bass 3: Der Basssound wurde mit dem Twin 2 Synthesizer von Fabfilter erzeugt. Ein LFO-gesteuerter High-Cut-Filter sorgt für etwas Bewegung.

Synth Lead 3: Der Sound wurde mit Massive von Native Instruments synthetisiert, dieser zeichnet sich durch seine Veränderung aus: Zunächst erklingt die gespielte Melodie trocken und in mono, danach kommen sukzessive Effekte hinzu: Der Sound wird mittels Chorus dekorreliert, Reverb und Delay verstärken den räumlichen Effekt. Hinsichtlich der Korrelation fluktuiert der Sound am Ende zwischen 0,25 und 0,5.

Noise Oszillation: Nach ein paar Takten setzt ein in geraden Sechzehntelnoten spielender Sound ein, der aus rosa Rauschen mit einem Granularsynthesizer erzeugt wurde. Der Klang wird zunehmend lauter und hat einen Korrelationswert von 0.

Kompositorisch fungiert der Formteil, wie schon erwähnt, als Aufbau zum letzten Abschnitt. Die erklingende Melodie des Synth Lead 3 bildet dabei den Kern. Diese Melodie wurde schon im Formteil B als “Synth Hintergrund” angedeutet, derartige musikalischen Verbindungen sollen für die erwähnte kompositorische Kohärenz sorgen.
Teil E weist am Ende einen Korrelationswert von unter 0,25 auf, wobei der Wert in den unteren Mitten gegen 0 geht.

6.1.7 Teil F (3:33 - 4:30)

Nachdem der vorhergehende Teil abrupt endet, spielt am Beginn des letzten Formteils F die Hauptmelodie von Teil E. Diese erklingt solo und in einem trockenen Mono-Sound, so soll erneut eine Kontrastwirkung erzielt werden. Nach einem Takt setzen die restlichen Klangkomponenten ein.
Kompositorisch und klangtechnisch ist Teil F, abgesehen von ein paar Details, gleich wie Teil B. Die erwähnte Melodie von Teil E spielt auch über den gesamten letzten Teil und in der zweiten Hälfte von Teil F kommen noch vier Gesangsspuren hinzu. Die Gesangsstimmen wurden mittels Panning aufgeteilt und das Seitensignal der Summe wurde bei 5,5 kHz etwas angehoben. Zusätzlich wurde ein Reverb verwendet, um der Stimme einen “träumerischen” Charakter zu geben.
Das Stück endet abrupt mit einem Schaltergeräusch.

6.2 Room Switch Kurzversion

Da die Hauptkomposition mit 4:32 min. für Demonstrationen innerhalb von Vorträgen oder anderen zeitlich knapp bemessenen Gelegenheiten etwas zu lang ist, wurde eine alternative Version des Stücks angefertigt. Wie bei der normalen Version ist auch diese in sechs Formteile gegliedert, jedoch wurde versucht die Formteile zeitlich so herunter zu brechen, dass nur die wichtigsten Aspekte erhalten bleiben. Die Kurzversion ist daher mit 1:28 min. deutlich kürzer als das Hauptstück und bietet

trotzdem die wesentlichen Höreindrücke.

7. Fazit

Bei der Produktion des Stücks “Room Switch” wurden die Erkenntnisse die aus der vorausgehenden Analysearbeit gewonnen wurden praktisch umgesetzt. Dabei konnte bestätigt werden, dass die aus der Analyse entstammten Kriterien, wie Dekorrelation, Atmosphäre, Reverb, etc. für eine immersive Audioproduktion zur Wiedergabe über den Echo Studio essentiell sind:
Es konnte beobachtet werden, dass Dekorrelation abhängig von der Art des Audiosignals unterschiedliche Auswirkungen auf die Wiedergabe hat: Atmosphärische Sounds profitieren im Sinne des LEVs von einem niedrigen Korrelationswert, konkretere Klänge können durch Phasenverschiebung vom Smartspeaker externalisiert werden. Letzteres wurde in Formteil B bei der Snare-Drum und Hihat angewandt.
Spektral dichte atmosphärische Klänge funktionieren über den Echo Studio besonders gut, weshalb diese eine wichtige Rolle in der präsentierten Komposition einnahmen. Das zusätzliche binaurale Panning solcher Sounds bietet die Möglichkeit diese grob im Raum zu platzieren.
Bewegungen konnten mit automatisierten Low-Cut-Filter und Flanger realisiert werden. Erstere wurden für Pad-Sounds verwendet, wodurch sich das Frequenzspektrum stetig ändert und es so zu unterschiedlichen Höheneindrücken kommt, da hohe Frequenzen aus psychoakustischer Sicht von oben wahrgenommen werden. Dieser Effekt konnte vor allem auch bei der verwendeten Helikopteraufnahme, die bei 2:53 min. zu hören ist, beobachtet werden.
Ein wichtiger Punkt der bei der Analyse noch nicht bedacht wurde, sondern erst bei der Produktion deutlich wurde, ist das Spiel mit Kontrasten. Räumliche Kontraste als Stilmittel in der Audioproduktion im Kontext von Stereo-Upmixing ist zur Verdeutlichung der Wirkungsweise dieser Technologie sehr hilfreich. Durch abrupte “Raumwechsel”, wie sie im vorgestellten Stück “Room Switch” stattfinden, können immersive Unterschiede deutlich wahrgenommen werden. Weiters zeigte sich hinsichtlich der Kontraste, dass es nicht nur auf den Wechsel von Mono und Stereo ankommt, sondern vor allem auf kompositorische Überlegungen. So wirken basslastige Sounds in mono deutlich einhüllender als höhenlastige Sounds in mono. Um also wirksame räumliche Kontraste zu gestalten, muss auch über die Soundwahl nachgedacht werden.

Das produzierte Stück stellt weiters die Frage nach Konventionen in den Raum. Formteil B wurde so gestaltet, dass möglichst alle Klangkomponenten vom Speaker externalisiert klingen und so einen maximalen Einhüllungsgrad erzielen. Dies hat

jedoch zur Folge, dass das Stück Korrelationswerte im negativen Bereich aufweist, wodurch es bei einer Wiedergabe in mono zu Auslöschungen kommen würde. Speziell bei kommerziellen Audioproduktionen spielt Monokompatibilität nach wie vor eine Rolle, denn obwohl Audio hauptsächlich in stereo konsumiert wird, gibt es noch immer viele Geräte und Situationen wo nur Mono eine Option ist. Dies ist zum Beispiel in vielen Discotheken, öffentlichen Bereichen wie Fußballstadien und Supermärkten der Fall, auch Radioübertragungen wechseln bei schlechten Empfang oft auf Mono. (vgl. delamar.de) Aber der/die Produzent*in kann und soll natürlich selbst entscheiden wie wichtig Monokompatibilität für das jeweilige Projekt ist.
Das Thema Mixing-Konventionen betrifft jedoch nicht nur die Monokompatibilität, sondern auch die Platzierung gewisser Sounds an gewisse Stellen im Stereofeld. So ist es üblich, dass die Kick- und die Snare-Drum, sowie der Gesang und der Bass mittig platziert werden. Natürlich obliegen diese Entscheidungen immer dem/der Produzentin, trotzdem ist es ein Punkt der bedacht werden sollte. Denn mit mehr verfügbarer Räumlichkeit, kommt auch das Verlangen diese zu nutzen, dies gilt noch mehr für 3D Audio. Als Produzent*in sollte man sich diesbezüglich jedenfalls Gedanken machen, auch wenn es auf den ersten Blick trivial erscheint.

Bei der Audioproduktion im Kontext von stereo-upmixfähigen Wiedergabesystemen muss grundsätzlich überlegt werden, ob der Audiocontent für ein bestimmtes Wiedergabesystem und nur für die Upmixfunktion gestaltet wird, oder ob der Content grundsätzlich als normale Stereoproduktion funktionieren soll und in einem Upmix-Szenario ebenfalls. Aus kommerzieller Sicht macht letzteres wohl mehr Sinn und ist letzten Endes auch das Ziel dieser Arbeit.
Anhand des Stücks “Room Switch” wurden beide Produktionsmöglichkeiten ausprobiert. Wie schon erwähnt, wurde in Formteil B das Konzept der Dekorrelation ausgereizt, was eine geringe Monokompatibilität zur Folge hatte. Formteil C hingegen wurde nach konventionelleren “Regeln” produziert und funktioniert auch als normaler Stereocontent der zudem monokompatibel ist.

Was lässt sich nun zum behandelten Thema hinsichtlich Audioproduktion im gesamten Spektrum sagen? Oder konkreter formuliert: Sollte in Zukunft auf das Thema Upmix bei der Audioproduktion Rücksicht genommen werden? Wir leben in einer Zeit wo Räumlichkeit und Immersion scheinbar immer wichtiger wird. Dies lässt sich anhand der Entwicklungen im 3D-Audiobereich beobachten, oder den auf Youtube sehr beliebten “8D-Audio”-Mixes, und nicht zuletzt anhand der Stereobreite der meisten modernen Popsongs. Auch die Verfügbarkeit der Upmix-Funktion im Echo Studio und ähnlichen Geräten kann als Indiz für diese Entwicklung gesehen werden. Im Laufe der vorliegenden Arbeit konnten einige Punkte zur Produktion im Kontext der Raumklangfunktion des Echo Studios formuliert werden. Die meisten dieser Punkte sind auch für normale Stereoproduktionen relevant und werden vermutlich von vielen Produzent*innen ohnehin bedacht. Man könnte also sagen,

dass der Zeitgeist in gewisser Weise schon Rücksicht auf stereo-upmixkompatible Produktionen nimmt. Im Zuge der vorliegenden Arbeit wurden gute immersive Ergebnisse erzielt, daher macht es aus der Sicht eines/einer Produzent*in durchaus Sinn diese Technologie zu bedenken, wenngleich es eher unwahrscheinlich ist, dass in Zukunft speziell für upmixfähige Geräte produziert wird. Doch hier kommt der oben erwähnte Punkt wieder ins Spiel, dass es ohnehin sinnvoller ist Audiocontent zu produzieren, der sowohl im Upmix-Kontext, als auch bei normaler Stereo-Wiedergabe funktioniert. Diesbezüglich sind die aus der Arbeit gewonnenen Erkenntnisse wegweisend und sollten prinzipiell bei der Produktion von Stereocontent bedacht werden.

Obwohl sich die Arbeit mit dem Fallbeispiel Echo Studio beschäftigt hat, wirken die gewonnen Erkenntnisse universell einsetzbar. Dies zu bestätigen indem man sich mit einigen weiteren stereo-upmixfähigen Geräten beschäftigt, kann als Ausblick für die weitere Forschung zum Thema gegeben werden.

8. Quellenverzeichnis

Amazon.de: Wie stellen vor: Echo Studio - Smarter High Fidelity-Lautsprecher mit 3D-Audio und Alexa. Letzter Zugriff 11.11.2020, https://www.amazon.de/amazon-echo-studio-smarter-high-fidelity-lautsprecher-mit-3d-audio-und-alexa/dp/B07NQDHC7S

Avni, Amir/Rafaely, Boaz: Interaural cross correlation and spatial correlation in a sound field represented by spherical harmonics. In: Ambisonics Symposium (2009)

Baskind, Alexis u.a.: Surround and 3D-Audio Production on Two-Channel and 2D-Multichannel Loudspeaker Setups. In: HAL archives-ouvertes (2015)

Choueiri, Edgar: Binaural Audio Through Loudspeakers. In: Immersive Sound. The art and science of binaural and multi-channel audio. Hg: Roginska, Agnieszka/Geluso, Paul (2017), S. 125-179

Cousins, M. P./Bleeck, S./Fazi, F. M.: The Effect of Inter-channel Cross-correlation Coefficient on Perceived Diffuseness. In: 4th International Conference on Spatial Audio (2017)

Delamar.de: Abmischen. Die Sache mit der Monokompatibilität. Letzter Zugriff 22.12.2020, https://www.delamar.de/mixing/abmischen-monokompatibilitaet-5610/

Elen, Richard: Ambisonics. The surround alternative. (2001)

Floros, Andreas/Tatlas, Nicolas-Alexander: Spatial enhancement for immersive audio applications. In: International Conference on Digital Signal Processing (2011)

Geluso, Paul: Stereo. In: Immersive Sound. The art and science of binaural and multi-channel audio. Hg: Roginska, Agnieszka/Geluso, Paul (2017), S. 63-87

He, Jianjun/Gan, Woon-Seng: Applying Primary Ambient Extraction for Immersive Spatial Audio Reproduction. In: Proceedings of APSIPA Annual Summit and Conference (2015), S. 1000-1009

He, Jianjun: Spatial audio reproduction using primary ambient extraction. (2016)

Hooley, Tony: Single box surround sound. In: Acoustical Science and Technology 27 (6) (2006), S. 354-360

Ibrahim, Karim M./Allam, Mahmoud: Primary-Ambient Source Separation for Upmixing to Surround Sound Systems. In: International Conference on Acoustics, Speech and Signal Processing (2018), S. 431-435

Jackson, Philip J. B. u.a.: Object-Based Audio Rendering. (2017)

Käsbach, Johannes: Characterizing apparent source width perception. (2016)

Kendall, Gary S.: The Decorrelation of Audio Signals and its Impact on Spatial Imagery. In: Computer Music Journal 19(4) (1995), S. 71-87

Kraft, Sebastian: Email: Betreff: Trennung von Direktsignal und Ambientsignal. Adresse: skraft@hsu-hh.de am 14.7.2020

Kraft, Sebastian/Zölzer, Udo: Stereo Signal Separation and Upmixing by Mid-Side Decomposition in the Frequency-Domain. In: Proc. Of the 18th Int. Conference on Digital Audio Effects (2015), S. 1-6

Lee, Hyunkook: Apparent source width and listener envelopment in relation to source-listener distance. In: AES 52nd International Conference (2013), S. 1-6

Pfanzagl-Cardone, Edwin: Signalkorrelation und Raumeindruck bei Stereo- und 5.1 Surround-Aufnahmen. (2010)

Pulkki, Ville: Directional audio coding in spatial sound reproduction and stereo upmixing. In: AES 28th International Conference (2006), S. 1-8

Pulkki, Ville/Delikaris-Manias, Symeon/Politis, Archontis: Parametric Time-Frequency Domain Spatial Audio. (2017)

Rampelmann, Klaus: Nachhallzeit, darf's auch ein bißchen mehr sein? (2015)

Roginska, Agnieszka: Binaural Audio Through Headphones. In: Immersive Sound. The art and science of binaural and multi-channel audio. Hg: Roginska, Agnieszka/Geluso, Paul (2017), S. 88-123

Roginska, Agnieszka/Geluso, Paul: Introduction. In: Immersive Sound. The art and science of binaural and multi-channel audio. Hg: Roginska, Agnieszka/Geluso, Paul (2017), S. 1-4

Rozenn, Nicol: Sound Field. In: Immersive Sound. The art and science of binaural and multi-channel audio. Hg: Roginska, Agnieszka/Geluso, Paul (2017), S. 276-310

Rumsey, Francis: Surround Sound. In: Immersive Sound. The art and science of binaural and multi-channel audio. Hg: Roginska, Agnieszka/Geluso, Paul (2017), S. 180-220

Sato, Shin-ichi/Ando, Yoichi: Apparent source width (ASW) of complex noises in relation to the interaural cross-correlation function. In: 17th International Congress on Acoustics (2001)

Sengpiel, Eberhard: Transaural Stereo. Kunstkopf-Stereofonie über Lautsprecher. (1998)

Sengpielaudio: Berechnen der Nachhallzeit. Letzter Zugriff 5.1.2021, http://www.sengpielaudio.com/Rechner-RT60.htm

Sharma, Gerriet K.: Komponieren mit skulpturalen Klangphänomenen in der Computermusik. (2016)

Sontacchi, Alois/Höldrich, Robert: Schallfeldreproduktion durch ein verbessertes Holophonoie-Ambisonic System. (2000)

Tsingos, Nicolas: Object-Based Audio. In: Immersive Sound. The art and science of binaural and multi-channel audio. Hg: Roginska, Agnieszka/Geluso, Paul (2017), S. 244-275

Wenzel, Elisabeth M./Begault, Durand R./Godfroy-Cooper,Martine: Perception of Spatial Sound. In: Immersive Sound. The art and science of binaural and multi-channel audio. Hg: Roginska, Agnieszka/Geluso, Paul (2017), S. 5-39

Zotter, Franz/Frank, Matthias: Ambisonics. A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality. (2019)

9. Abbildungsverzeichnis

3D-Audio und Alexa. Letzter Zugriff 11.11.2020, https://www.amazon.de/amazon-echo-studio-smarter-high-fidelity-lautsprecher-mit-3d-audio-und-alexa/dp/B07NQDHC7S

Abb. 6 - Abb. 30: Wurden selbst erzeugt.

Link zu den Audioproduktionen:

https://drive.google.com/drive/folders/1kzdljnEf8zIB3HPk3aK5KKb1PdiSaSXU