Exposé zur Dissertation

Verfahren zur Parametrisierung von Schallquellen und deren Interaktion mit dem Raum in 3D Aufnahmen

Daniel Rudrich

Juni 2017

1 Einleitung

3D Aufnahmen von akustischen Szenen bestehen nicht nur aus den vorhandenen Direktsignalen von Schallquellen sondern auch aus deren Interaktion mit dem umschließenden Raum. Am Hörort treffen daher von einer Schallquelle herrührend mehrere Schallereignisse ein. Dazu zählen neben dem Direktsignal auch Reflexionen an Objekten und Raumbegrenzungen sowie Ereignisse, die durch Abschattungs- und Beugungseffekte entstehen.

Tools zur Bearbeitung von Ambisonischen Signalen [1] erlauben es, in solchen Aufnahmen unerwünschte Schallquellen zu entfernen, indem die Signale aus den Richtungen dieser Quellen ausgeblendet werden. Dabei berücksichtigt man jedoch lediglich den Direktschall der zu unterdrückenden Quellen. Die Reflexionen von begrenzenden Raumflächen treffen generell aus anderen Richtungen und zu anderen Zeitpunkten am Hörort ein. Dies führt dazu, dass zwar der Direktschall gedämpft wird, die Wandreflexionen und auch der diffuse Nachhall in der Aufnahme bestehen bleiben. Zudem könnten dadurch auch ungewollt Reflexionen anderer Quellen ausgeblendet werden.

In dieser Dissertation wird das Ziel verfolgt, in einer bestehenden 3D-Aufnahme unter Zuhilfenahme einer gegebenen Quellrichtungsvorgabe des Anwenders bzw. der Anwenderin die vollständige Rauminformation (Reflexionen, später Nachhall) zu parametrisieren und damit die Schallquelle als eigenständiges Objekt in der akustischen Szene greifbar zu machen. Das bedeutet, Quellen samt ihrer Interaktion mit dem Raum sollen getrennt von einander beschrieben werden. In erster Instanz werden einzelne diskrete Reflexionen behandelt, die mit Hilfe des Spiegelquellen-Modells (siehe Abschnitt 2) beschrieben werden können. Es sollen alle zeitlich und/oder räumlich trennbaren Reflexionen erfasst werden. Der verbleibende wahrnehmbare Nachhall der Quelle wird zudem statistisch beschrieben.

Die Parametrisierung soll einen nachträglichen manipulativen Eingriff in die aufgenommene Szene ermöglichen. Darunter fällt unter anderem das oben genannte Entfernen einer Quelle. Weitere denkbare Anwendungsmöglichkeiten sind in Abschnitt 4 aufgezeigt.

2 Spiegelquellen-Modell und Raumreflexionen

Reflexionen an Raumbegrenzungen lassen sich mit Hilfe des Spiegelquellen-Modells [2] darstellen und berechnen. Dabei wird die Quelle an den Raumbegrenzungsflächen gespiegelt. Es entstehen Spiegelquellen erster Ordnung. Durch weitere Spiegelungen entstehen Spiegelquellen höherer Ordnung. Abbildung 1 zeigt exemplarisch drei Spiegelquellen einer Schallquelle im Raum. Die Spiegelquellen werden als eigenständige Quellen betrachtet. Die tatsächlichen Reflexionspfade (dicke Linien) stimmen bezüglich Länge und Richtung mit denen der Spiegelquellen (dünne Linien) überein. Es ist zu erkennen, dass Spiegelquellen - und damit die Raumreflexionen - sich hinsichtlich Richtung und Signallaufzeit von der ursprünglichen Quelle unterscheiden.

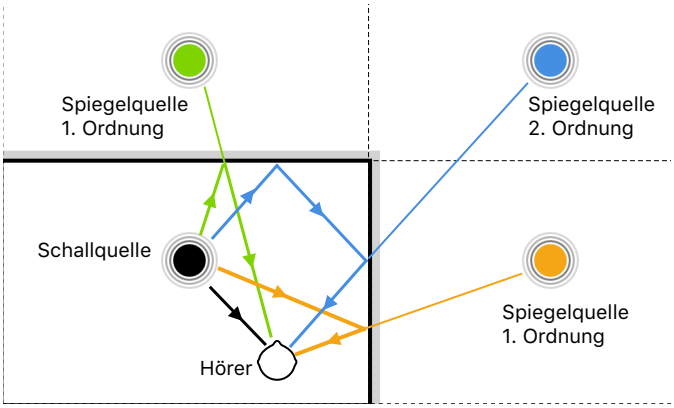

Abbildung 1: Exemplarische Darstellung des Spiegelquellen-Modells. Die Schallquelle (schwarz) wird an den Wänden gespiegelt (grün, orange). Eine Spiegelung an zwei Wänden führt zu Spiegelquellen 2. Ordnung (blau).

Durch die längere Wegstrecke und Absorption der Begrenzungsflächen treffen sie zudem mit niedrigerem Pegel[1] und spektral gefiltert am Hörort ein. Für bewegte Quellen entstehen zusätzlich noch unterschiedliche Doppler-Effekte [3] für Originalquelle und deren Reflexionen. Die Raumreflexionen tragen zur Raum- und Distanzwahrnehmung bei, werden jedoch durch den Haas-Effekt [4] nicht als separate Schallquellen wahrgenommen. Dies führt dazu, das sie sich nicht separat lokalisieren lassen und man somit ein Plugin zur Richtungsdämpfung nicht dementsprechend einstellen kann. Je nach Lage der Schallquellen im Raum und dessen Beschaffenheit ergeben sich unterschiedliche Reflexionsmuster, wobei aus ein und derselben Raumrichtung mehrere Reflexionen als auch Direktschall der einzelnen Ereignisse kommen können.

[1] Ausnahme sind stark gerichtete Quellen, bei denen die Reflexionen einen höheren Pegel aufweisen als das Direktsignal.

3 Parameterschätzung

Die zu schätzenden Parameter der Schallereignisse in der Aufnahme sind deren DOA (direction of arrival, Einfallsrichtung) und TDOA (time difference of arrival, relative Ankunftszeit). In Abhängigkeit von der Raumbeschaffenheit und Lage einer Schallquelle (DOA und TDOA) sowie deren Charakteristik (zeitlich und spektrale Zusammensetzung) soll eine Gruppierung der Schallereignisse durchgeführt werden. Somit wird eine Schallquelle beschrieben durch ein Primärereignis und den ihr zuordenbaren Reflexionen.

Sun et al. [5] geben einen Vergleich verschiedener Methoden zur Schätzung der DOA von Direktsignal und Reflexionen in Ambisonischen Aufnahmen. Dabei liefern die EB-MUSIC Methode (Eigenbeam Multiple Signal Classification) [6] und der EB-MVDR (Eigenbeam Minimum Variance Distortionless Response) Beamformer [7] die besten Ergebnisse. Erstere benötigt jedoch a-priori die Anzahl der Schallereignisse. Eine Zuweisung von Reflexionen zu einer Quelle wird jedoch in beiden Fällen nicht durchgeführt.

Einen vielversprechenden Ansatz zeigen Mabande et al. [8] zur Schätzung von DOA und TDOA von Reflexionen. Dabei wird ebenso der oben genannte EB-MVDR Beamformer eingesetzt, welcher statistisch optimiert das Signal aus der gewünschten Richtung unverzerrt wiedergibt. Unter Vorgabe des Direktsignals, der Position der Quelle im Raum und den errechneten Daten erfolgt zusätzlich eine Raumgeometrieschätzung und die Ermittlung der Reflexionskoeffizienten der Raumbegrenzungsflächen. Eine Zuweisung von Reflexionen zu einer Quelle geschieht dort allerdings nur indirekt über die TDOA-Schätzung mittels Kreuzkorrelation.

Um die Vorgabe der Gruppierung von Schallereignissen zu ermöglichen, müssen hier die zeitlich-spektralen Zusammenhänge berücksichtigt werden. Eine Alternative zu den oben genannten Ansätzen stellt die konsequente Erweiterung des von Smaragdis und Raj [9] präsentierten Verfahrens um das räumliche Attribut dar. Ausgehend von [10], die eine automatische Transkription von musikalischen Ereignissen anhand spektraler Eigenschaften und zuordenbaren Zeitpunkten durchführt, wird in [11] ein zeitlich-spektraler Verbund (z.B. Anschlag einer Note mit Ausklang) als Muster durch Entfaltung gesucht. Mit der *Shift-Invariant Probalistic Latent Component Analysis (PLCA)* [9] wird das zu suchende Muster nicht als konkrete Realisation sondern durch statistische Verteilungen beschrieben. Zudem kann der Suchraum auf beliebige Dimensionen erweitert werden und somit neben Zeit und Frequenz auch die räumlichen Dimensionen miteinbeziehen.

Die sowohl zeitlichen als auch spektralen Muster von Reflexionen korrelieren stark mit denen der Quelle und können sich bedingt durch Raum und Dopplereffekt in den Dimensionen Zeit, Richtung und Frequenz unterscheiden. Die Suche mittels *Shift-Invariant*

PLCA stellt daher ein vielversprechendes Verfahren zur Suche von Reflexionen dar, welches zudem die Schallereignisse entsprechend gruppiert. Durch die Verwendung einer *Constant-Q Transformation* [12] macht sich der Dopplereffekt als eine lineare Verschiebung des Musters im Frequenzbereich bemerkbar, was das Wiederauffinden des Musters erleichtert.

Pessentheiner [13] zeigt, dass durch Analyse harmonischer Komponenten - ergänzend zur Ortsinformation - gleichzeitig aktive Schallereignisse verbessert lokalisiert und charakterisiert werden können. So lassen sich beispielsweise zwei Sprecher mit unterschiedlichen Grundfrequenzen aber gleicher Richtung identifizieren, da deren Obertöne nicht zusammenfallen.

Durch die Vorgabe der Richtung der zu analysierenden Quellen durch die Anwenderin bzw. den Anwender muss keine Blindschätzung erfolgen. Jedoch liegt das Direktsignal nicht isoliert vor, sondern kann sowohl eigene Reflexionen und/oder die anderer Quellen, als auch einen Diffusanteil beinhalten. Durch Einbezug der Reflexionen lässt sich mit Rake Receiver Methoden wie [14, 15] der Signal-Stör-Abstand (SIR, signal to interference ratio) des Direktsignals erhöhen. Zu Störern zählen alle unerwünschten Komponenten wie Reflexionen, Diffushall oder Signalkomponenten anderer Quellen. Durch diese Verbesserung der Extraktion des Quellsignals kann wiederum die Parameterschätzung der Reflexionen rekursiv optimiert werden. Zusätzlich kann die Verwendung von Post-Filtern die Gewinnung des Direktsignals begünstigen, wie in [16] gezeigt.

Werden alle Schallereignisse der Aufnahme ausreichend genau beschrieben und charakterisiert, so können diese zu entsprechenden verursachenden Schallquellen zusammengefasst werden. Für jede dieser Quellen lassen sich Quellsignal und zugehörige Raumimpulsantwort definieren. Damit besteht die Möglichkeit, auf einzelne Schallquellen und dazu zuordenbare Schallereignisse unmittelbar zuzugreifen.

4 Mögliche Anwendungen

Anwendungen des Verfahrens befinden sich vor allem in der Postproduktion von Ambisonischen Aufnahmen. Durch die resultierende Greifbarkeit der einzelnen beitragenden Quellen ist es möglich, diese unabhängig voneinander zu bearbeiten. Einerseits können einzelne Quellen hervorgehoben oder abgesenkt werden, unabhängig davon auch deren Raumreflexionen. Dadurch lassen sich zum Beispiel perzeptive Effekte wie die wahrgenommende Quellbreite (apparent source width, ASW) oder die Umhüllung (listener envelopment, LEV)[17] kontrolliert steuern.

Wird der Pegel von Direktschall und Reflexionen einer Quelle ausgeblendet, werden alle Beiträge einer Schallquelle zur Aufnahme entfernt. Darüberhinaus kann an diese Quellpositionen eine andere Quelle eingesetzt werden, die vorher nicht im Aufnahmeraum vorhanden war.

Durch die Hinzunahme eines plausiblen Raummodells können die vorhanden Quellen im Raum bewegt werden. So lassen sich zwei Quellen räumlich stärker separieren, oder eine Quelle näher an den Hörort schieben und sie somit präsenter machen. Beides unter Beibehaltung des physikalischen Raumes.

Der momentan mögliche Eingriff, der durch die Verfügbarkeit der räumlichen Information gegeben ist [1], stellt eine enorme Verbesserung gegenüber Stereo-Produktionen dar. Durch das in dieser Dissertation angestrebte Verfahren wird der Eingriff grundsätzlich subtiler und hält der physikalischen Plausibilität stand.

5 Vorgehensweise, mögliche Schwierigkeiten, Evaluation

Beginnend mit einer simulierten Szene mit Raumsimulationen mittels der Spiegelquellen-Methode[2] soll ein Algorithmus entwickelt werden, der die oben beschriebenen Parameter schätzt. Da die Grundwahrheiten durch die Simulation bekannt sind, lassen sich die verschiedenen Ansätze abwägen und optimieren. Neben der frei wählbaren Signale (z.B. rosa Rauschen, Sprache, geeignete Messsignale) lassen sich dabei auch kritische Situationen herstellen, mit mehreren gleichzeitig aktiven Quellen, überlappenden Schallereignisrichtungen, ähnlichen harmonischen Komponenten der einzelnen Signale oder bewegte Quellen.

Darauf aufbauend sollen kontrollierte reale Aufnahmen erstellt und der Algorithmus dorthin optimiert werden. Diese lassen sich beispielsweise mit im Raum verteilten Lautsprechern realisieren, wobei Signal und Raumgeometrie bekannt bzw. messbar sind.

Um den Einsatz des Algorithmus unter realen Bedingungen zu untersuchen, werden komplexere Szenen ohne kontrollierbare Eigenschaften verwendet. Diese sind beispielsweise Konzertaufnahmen oder Aufnahmen in urbanen Räumen.

Reale Ambisonische Mikrofone verfügen generell nicht über eine perfekte Enkodierung der Signale in Kugelflächenfunktionen und verfügen nur über eine endliche Richtungsauflösung. Deswegen tritt bei hohen Frequenzen *Spatial Aliasing*[18] auf, es kommt zu räumlichen Verzerrungen und Abbildungsfehlern. Zudem kann bei nahen Quellen in den

[2] Geeignete frei verfügbare/kommerzielle Produkte zur Raumsimulation sind beispielsweise McRoomSim, EASE oder CATT-Acoustics.

tiefen Frequenzen nicht von einer ebenen Welle ausgegangen werden, was zusätzlich Enkodierungsfehler erzeugt. Diese Enkodierungsartefakte stellen zusammen mit räumlich, zeitlich und spektral dicht gepackten Schallereignissen die größten Hürden für die Parametrisierung dar.

Die Praxistauglichkeit wird am Ende primär mit Hörversuchen attestiert, indem untersucht wird, inwieweit Artefakte in der Signalverarbeitung kaschiert werden müssen, damit diese nicht wahrnehmbar sind. Die zu untersuchenden Anwendungen sind:

- Ausblenden einer Quelle,
- Ersetzen eines Quellsignals mit einem szenenfremden Signals,
- Steuerung von ASW und LEV durch Pegeländerung der Reflexionen.

Daraus lassen sich folgende Fragestellungen ableiten:

- Treten Artefakte/Klangfärbungen der Schallquelle(n), des Raumes auf?
- Ist der Raum bzw. die Position der Schallquelle im Raum plausibel?
- Ist bei Ausblendung einer Schallquelle diese oder deren Reflexionen noch hörbar?
- Lassen sich Parameter wie ASW und LEV direkt kontrollieren?

Literatur

[1] M. Kronlachner, "Plug-in suite for mastering the production and playback in sorround sound ambisonics," in *Gold Award at AES Student Design Competition*, Berlin, Apr. 2014.

[2] H. Kuttruff, *Room Acoustics*. Crc Press, 2016.

[3] C. Doppler, *Über das farbige Licht der Doppelsterne und einiger anderer Gestirne des Himmels*. Verlag der Königl. Böhm. Gesellschaft der Wissenschaften, 1842.

[4] H. Haas, "Über den Einfluß eines Einfachechos auf die Hörsamkeit von Sprache," *Acta Acustica united with Acustica*, vol. 1, no. 2, pp. 49–58, 1951.

[5] H. Sun, E. Mabande, K. Kowalczyk, und W. Kellermann, "Localization of distinct reflections in rooms using spherical microphone array eigenbeam processing," *The Journal of the Acoustical Society of America*, vol. 131, no. 4, pp. 2828–2840, Apr. 2012.

[6] B. Rafaely, Y. Peled, M. Agmon, D. Khaykin, und E. Fisher, "Spherical Microphone Array Beamforming," in *Speech Processing in Modern Communication*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 281–305.

[7] H. Sun, E. Mabande, K. Kowalczyk, und W. Kellermann, "Joint DOA and TDOA estimation for 3D localization of reflective surfaces using eigenbeam mvdr and spherical microphone arrays," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2011, pp. 113–116.

[8] E. Mabande, K. Kowalczyk, H. Sun, und W. Kellermann, "Room geometry inference based on spherical microphone array eigenbeam processing," *The Journal of the Acoustical Society of America*, vol. 134, no. 4, pp. 2773–2789, Oct. 2013.

[9] P. Smaragdis und B. Raj, "Shift-Invariant Probabilistic Latent Component Analysis," *Journal of Machine Learning Research*, pp. 1–29, Dec. 2007.

[10] P. Smaragdis und J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Oct 2003, pp. 177–180.

[11] P. Smaragdis, "Non-negative Matrix Factor Deconvolution; Extraction of Multiple Sound Sources from Monophonic Inputs," in *5th International Conference on Independent Component Analysis and Blind Signal Separation, Grenada, Spain*, 2004.

[12] J. C. Brown, "Calculation of a constant Q spectral transform," *The Journal of the Acoustical Society of America*, vol. 89, no. 1, pp. 425–434, 1991.

[13] H. Pessentheiner, "Localization, Characterization, and Tracking of Harmonic Sources," Ph.D. dissertation, Graz University of Technology, Jan. 2017.

[14] I. Dokmanic, R. Scheibler, und M. Vetterli, "Raking the Cocktail Party," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 5, pp. 825–836, Jul. 2015.

[15] H. A. Javed, A. H. Moore, und P. A. Naylor, "Spherical harmonic rake receivers for dereverberation," in *2016 IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2016, pp. 1–5.

[16] S. Delikaris-Manias und V. Pulkki, "Cross Pattern Coherence Algorithm for Spatial Filtering Applications Utilizing Microphone Arrays," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 11, pp. 2356–2367, 2013.

[17] J. S. Bradley und G. A. Soulodre, "The influence of late arriving energy on spatial impression," *The Journal of the Acoustical Society of America*, vol. 97, no. 4, pp. 2263–2271, Apr. 1995.

[18] B. Rafaely, B. Weiss, und E. Bachmat, "Spatial Aliasing in Spherical Microphone Arrays," *IEEE Transactions on Signal Processing*, vol. 55, no. 3, pp. 1003–1010, Feb. 2007.