Dynamische binaurale Raumsynthese in der 3D-Tonwiedergabe – eine Untersuchung zur Qualität verschiedener virtueller Lautsprecherkonfigurationen

Jan-Hendrik Hanschke[1], Felix Fleischmann[1], Jan Plogsties[1] und Simone Füg[1]

[1] *Fraunhofer-Institut für Integrierte Schaltungen, 91058 Erlangen, Deutschland, Email: janhendrikhanschke@gmail.com*

Einleitung

Die Erweiterung von Audio-Wiedergabesystemen um eine dritte Dimension, also die zusätzliche Verwendung elevierter Lautsprecher, hat der Audiobranche in den vergangenen Jahren viel Anlass für Forschung und neue Entwicklungen gegeben. Die Installation einer aus vielen Lautsprechern bestehenden Wiedergabekonfiguration kann aber sehr aufwändig sein. Um auch ohne diesen Aufwand ein dreidimensionales Klangbild präsentieren zu können, kann die Technologie der Binauralsynthese verwendet werden. Dieses Verfahren ermöglicht es beispielsweise, eine Abhörumgebung mit vielen Lautsprechern virtuell über Kopfhörer zu erzeugen, ohne einen physikalischen Aufbau der Lautsprecher zu benötigen.

Dafür werden große Datensätze von binauralen Impulsantworten benötigt. Für die daraus resultierenden umfangreichen und langen Messreihen bieten sich grundsätzlich Kunstköpfe an. Individuelle anatomische Abweichungen in Kopf- und Außenohr-Form des Kunstkopfes äußern sich dabei aber unter anderem in Veränderungen des Timbre und Ungenauigkeiten der Lokalisation im Vergleich zum natürlichen Hören. Da aufgrund dieser Tatsache eine akustische Szene nur verzerrt wiedergegeben werden kann, soll untersucht werden, wie viele virtuelle Lautsprecher noch zu einer Verbesserung der perzeptiven Qualität führen.

Ziel dieser Arbeit ist es, die wahrgenommenen Qualitätsunterschiede verschiedener binaural synthetisierter Lautsprecherkonfigurationen zu evaluieren. Es wurde dafür in einem für 3D-Wiedergabe konzipierten Abhörraum eine nicht-individuelle Binauralsynthese umgesetzt. Ein Hörversuch wurde sowohl für binaurale Kopfhörer-, als auch für Lautsprecherwiedergabe durchgeführt, um die Qualität unterschiedlicher Lautsprecherkonfigurationen für beide Wiedergabeverfahren zu evaluieren und miteinander zu vergleichen.

Für die Untersuchung wurden zwei Hypothesen aufgestellt:

- Die bewertete Qualität steigt monoton zu Konfigurationen mit großer Lautsprecheranzahl an. Dabei tritt zunehmend eine Sättigung der empfundenen Qualität auf.
- Es existieren, begründet auf dem nicht-individuellen Impulsantworten-Datensatz, Unterschiede in der Wahrnehmung der Qualität zwischen Kopfhörer- und Lautsprecherwiedergabe.

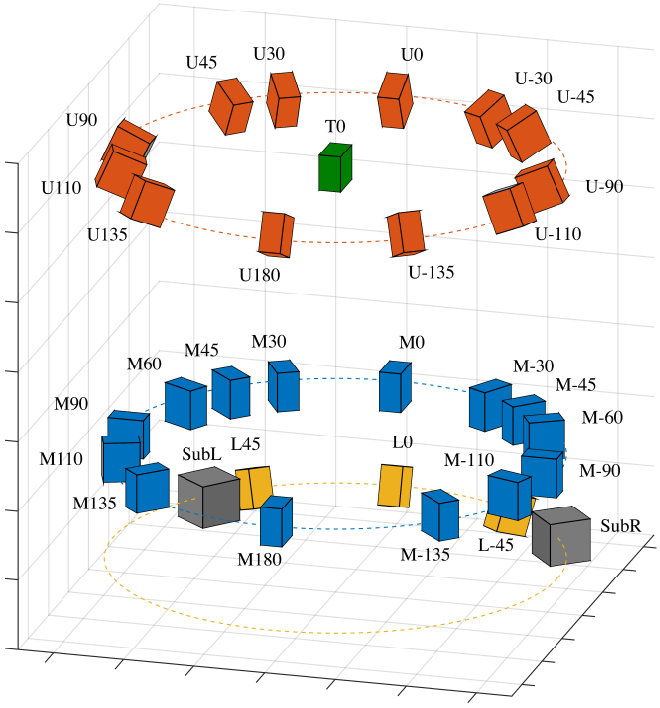

Abbildung 1: Das zur Messung der BRIR verwendete 32 Kanal Mess-Setup als Superset aller untersuchten Lautsprecherkonfigurationen. Die Entfernung der Lautsprecher zum Sweetspot betrug 2,8 m.

Auralisation einer Abhörumgebung

Die Binauralsynthese bezeichnet ein Verfahren, bei dem durch die Faltung eines Eingangssignals mit einer binauralen Raumimpulsantwort (BRIR) die akustischen Bedingungen „hörbar“ gemacht werden, unter denen die BRIR gemessen wurde. Dabei erfolgt im Falle der dynamischen Binauralsynthese in Echtzeit ein Austausch der entsprechenden Filterkoeffizienten für die aktuelle Kopfdrehung (Kopf-über-Torso-Orientierung) des Hörers. Als Grundlage für diese Auralisation wurden die nötigen BRIR-Messungen in einem für 3D-Wiedergabe konzipiertem, ITU-R BS.1116 konformen, Abhörraum [1] durchgeführt. Schallquellen waren 32 Lautsprecher an verschiedenen Positionen, mit denen das Abbilden von sämtlichen zu untersuchenden zwei- und dreidimensionalen Lautsprecherkonfigurationen möglich ist (siehe Abbildung 1).

Als Kunstkopf wurde ein am Fraunhofer IIS entwickelter Prototyp (Abbildung 2) verwendet. Da dieser mit einer Reihe von Motoren im Torso ausgestattet ist, ist es möglich, automatisch verschiedene Kopf-über-Torso-Orientierungen anzusteuern [2]. Insgesamt wurden in 101 dieser Orientierungen jeweils BRIRs aller

Abbildung 2: IIS Kunstkopf Prototyp mit Motoren im Torso zur Einstellung verschiedener Kopf-über-Torso-Orientierungen. DPA 4060 Mikrofonkapseln am geblockten Eingang des Gehörgangs.

Lautsprecher gemessen (=3232 Impulsantworten Paare). Für den in der Untersuchung verwendeten Kopfhörer (Stax SR-007 MKII) wurde zudem ein Invertierungsfilter erstellt, welcher am Messpunkt, dem geblockten Eingang des Gehörgangs, einen linearen Frequenzgang erzeugt, ohne jedoch individuelle schmalbandige Einbrüche im Frequenzgang zu invertieren. Das Faltungsprodukt von Filter und 0,7 s langen BRIRs wurde an eine Linux-Workstation übergeben, welche die Echtzeitberechnung der binauralen Signale mittels der Software *Sound Scape Renderer* [3] übernommen hat. Dabei wurden durchgängig alle 32 Kanäle berechnet, welchen ihr jeweiliges Eingangssignal über eine MADI-Schnittstelle zugespielt wurde. Als Headtracker kam ein Sensorboard mit integriertem Arduino-Microcontroller zum Einsatz (*Razor AHRS*).

Hörversuchs-Methodik

In einem Hörversuch sollten in der erstellten virtuellen Abhörumgebung verschiedene Lautsprecherkonfigurationen unterschiedlicher Komplexität miteinander verglichen werden.

Wie schon in Arbeiten von Kim et al. [4] und Silzle et al. [5] zur Untersuchung von 3D-Lautsprecherkonfigurationen kam auch hier die Hörversuchsmethodik nach ITU-R BS.1534-3 (MUSHRA) zum Einsatz. Probanden bewerteten in diesem Multi-Stimulus-Test mit versteckter Referenz und Ankersignal, wie sehr die empfundene Gesamtqualität der einzelnen Lautsprecherkonfigurationen von der Referenz abweicht. Auch in dieser Untersuchung stellte das NHK 22.2 Format durch seine hohe Lautsprecherdichte und die Möglichkeit, ein Klangbild in alle Richtungen rund um den Hörer zu erstellen, die Referenz dar. Es handelt sich dabei um die größte standardisierte Konfiguration. Die MUSHRA-Spezifikation sieht außerdem das Verwenden eines oder mehrerer Anker vor, welche eine deutliche Beeinträchtigung der Qualität beinhalten. Als *Low-Anchor* wurde hier eine stereofone Wiedergabe gewählt. Umhüllung, Raumwahrnehmung und Lokalisation weichen in diesem Fall deutlich von den dreidimensionalen Anordnungen ab.

Der eigentliche Bewertungsprozess erfolgte mittels einer kontinuierlichen Qualitätsskala, welche eine ganzzahlige Vergabe von 0 bis 100 Punkten erlaubt und in fünf Intervallen mit Adjektiven von *Bad* bis *Excellent* beschriftet ist. Die Probanden wurden instruiert die Gesamtqualität (*Overall Sound Quality*) zu bewerten, welche jegliche wahrgenommenen Unterschiede zwischen Referenz und System umfassen sollte. Dabei wurden einige Attribute und deren Beschreibung zur Bestimmung in der schriftlichen Hörversuchsanleitung angegeben: Umhüllung, Räumlichkeit, Realismus & Natürlichkeit, Lokalisation sowie Klangfarbe.

Lautsprecherkonfigurationen: Neben der 22.2 *Referenz* und dem Stereo *Low-Anchor* wurden im Hörversuch vier weitere Konfigurationen verglichen. Durch das Verwenden von zwei zwölf-kanaligen Konfigurationen (Auro 11.1 und 7.1+4; die Zahl hinter dem + bezeichnet die Anzahl Höhenlautsprecher) sollte der Einfluss einer unterschiedlichen Anordnung der selben Lautsprecheranzahl auf die Qualität überprüft werden. 5.1+2 stellte die kleinste standardisierte 3D-Konfiguration dar. Des weiteren wurde ein 5.1 Surround-Setup nach ITU-R BS.775-3 ausgewählt.

Stimuli: Die im Hörversuch verwendeten Stimuli sind ursprünglich für die 22.2 Konfiguration produziert worden und wurden für den Hörversuch mit dem im MPEG-H Standard festgelegten aktiven Downmixalgorithmus mit *phase-alignment* auf die kleineren Formate herunter gemischt [6]. Da die Ergebnisse des Hörversuchs von der Stimuli-Auswahl abhängig sind, wurde auf ein möglichst breites Spektrum von Genres, Anzahl der Quellen und Diffus- vs. Direktanteil geachtet. Insgesamt wurden sieben verschiedene Stimuli aus den Bereichen Filmton und Musik verwendet.

Ablauf: Um Vergleichswerte zu erhalten, wurde der Hörversuch in identischer Weise sowohl über reale Lautsprecher, als auch die Auralisation durchgeführt. Damit dabei möglichst viele Variablen (inkl. visueller Cues) gleich blieben, fanden beide Versuche in dem für die Messungen verwendeten Abhörraum statt. Insgesamt nahmen 20 Probanden an dem Versuch teil, die Erfahrungen mit Hörversuchen und/oder 3D-Tonwiedergabe hatten. Um Lerneffekte auszuschließen wurden die Probanden in zwei Gruppen aufgeteilt, von denen die eine mit dem Kopfhörer- und die andere mit dem Lautsprecherversuch begann.

Ergebnisse und Diskussion

Entsprechend der in der MUSHRA-Spezifikation vorgegebenen Kriterien für eine Nachselektion wurden zwei Probanden aus der Ergebnisauswertung ausgeschlossen. Anschließend wurden Mittelwert und 95%-

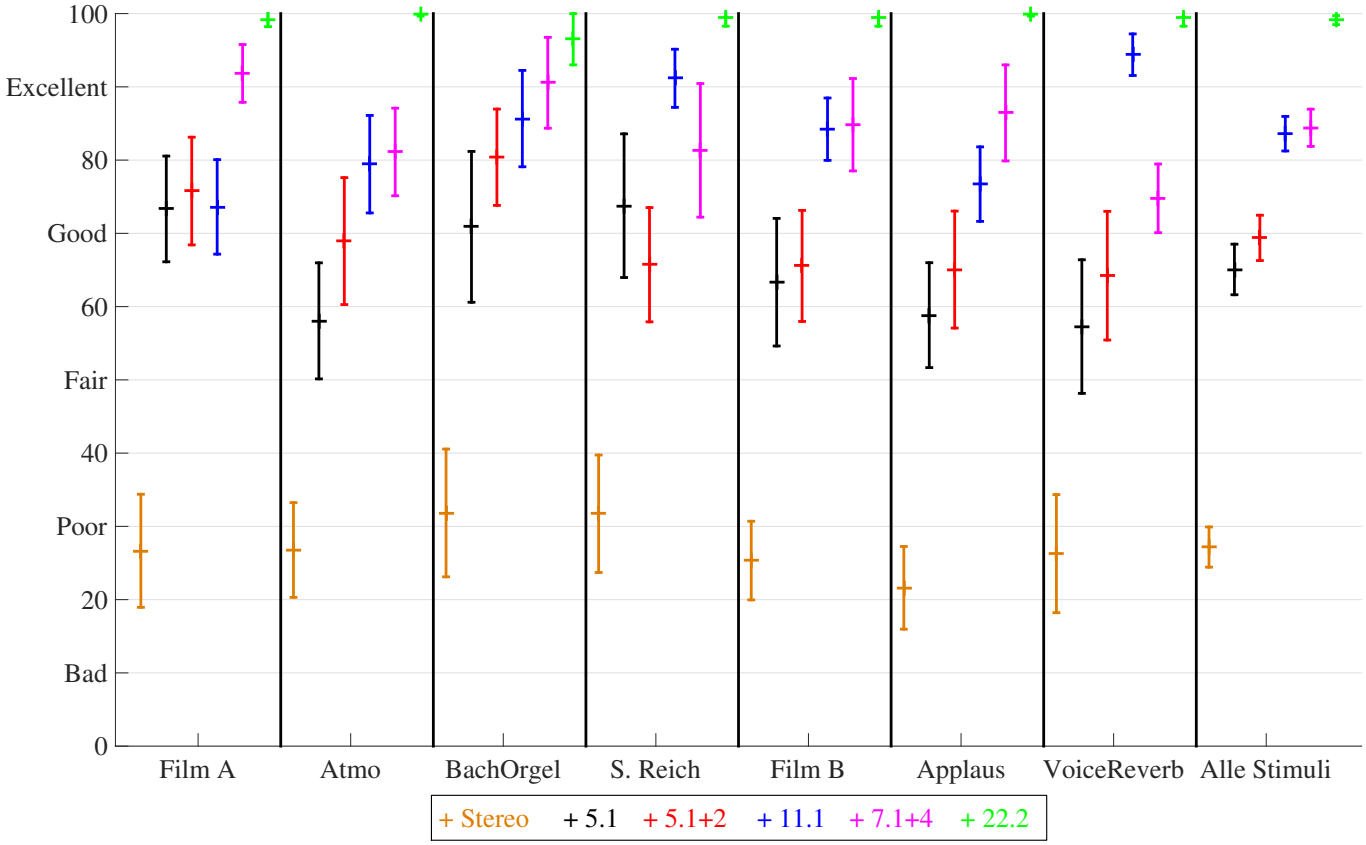

Abbildung 3: Ergebnisse des Hörversuchs - Binaural: Darstellung der Mittelwerte und 95%-Konfidenzintervalle von 18 Probanden. Bewertet wurden 6 Lautsprecherkonfigurationen und 7 Stimuli in einem MUSHRA-Testdesign.

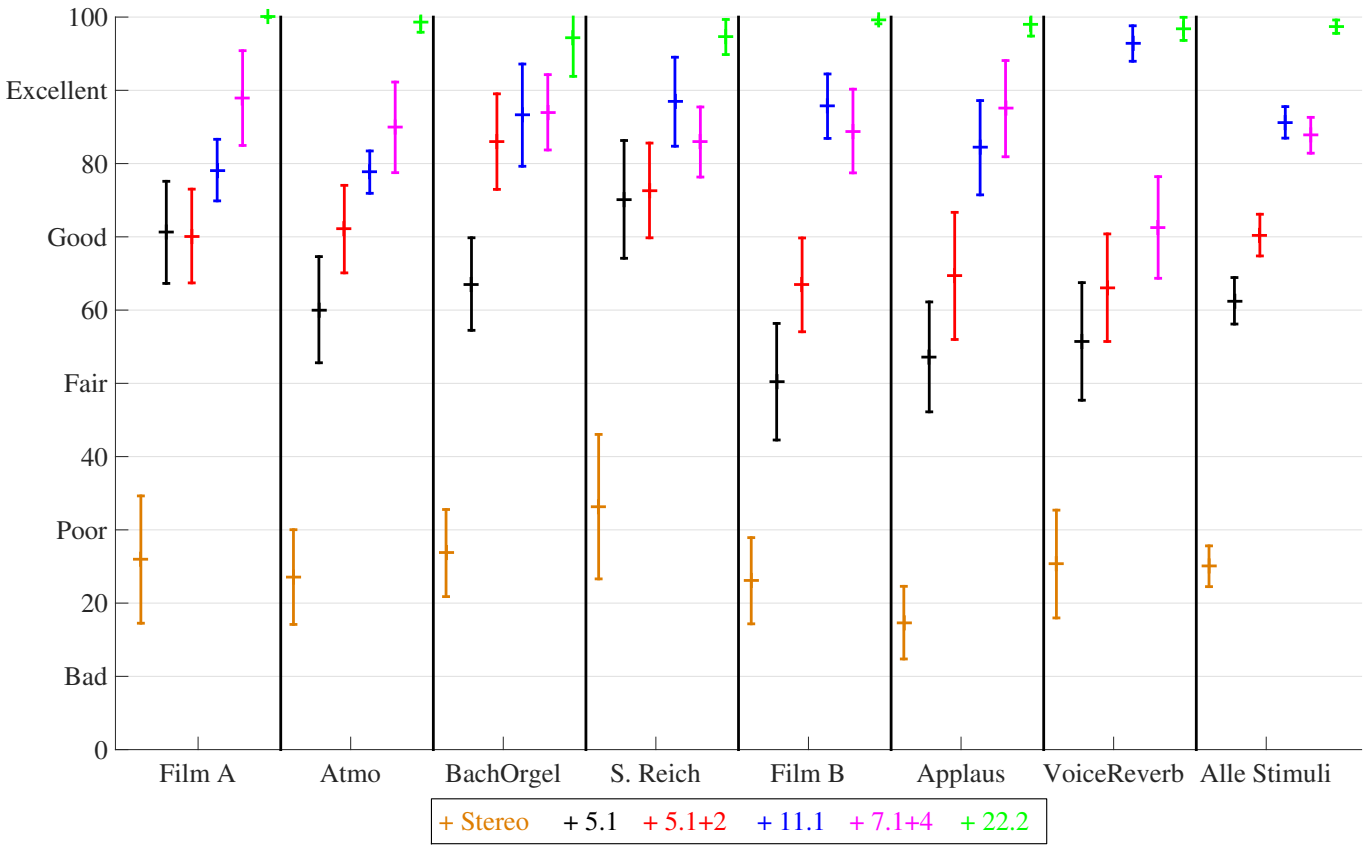

Abbildung 4: Ergebnisse des Hörversuchs - Lautsprecher: Darstellung der Mittelwerte und 95%-Konfidenzintervalle von 18 Probanden. Bewertet wurden 6 Lautsprecherkonfigurationen und 7 Stimuli in einem MUSHRA-Testdesign.

Konfidenzintervall aller Probanden und Stimuli dargestellt (Abbildung 3 und 4). Der Vergleich der beiden Ergebnisplots zeigt bereits die große Ähnlichkeit in den Ergebnissen von Lautsprecher- und binauraler Wiedergabe. Besonders bei der Analyse einzelner Stimuli fällt auf, dass die versteckte Referenz nicht immer eindeutig erkannt wurde. Die Auswertung der Einzelergebnisse der Probanden zeigt, dass nur die Hälfte der Teilnehmer über alle Stimuli hinweg in der Lage war, die Referenz einwandfrei von den Konfigurationen mit zwölf Lautsprechern zu unterscheiden. Dies lässt auf eine sehr große Ähnlichkeit dieser Lautsprecherkonfigurationen schließen. Im Mittel ist die 22.2-Referenz jedoch siginifikant besser bewertet worden als die anderen Anordnungen und war damit von diesen unterscheidbar. Die beiden auf zwölf Lautsprechern basierenden Konfigurationen unterscheiden sich zwar bei bestimmten Stimuli signifikant, wenn etwa an einer bestimmten Quellposition keine Phantomquelle, sondern ein diskretes (virtuelles) Lautsprechersignal wiedergegeben wird. Im Mittel über alle Stimuli ist allerdings kein signifikanter Unterschied erkennbar.

Die in der ersten Hypothese formulierte Annahme der Qualitätssättigung lässt sich an den in Abbildung 5 dargestellten Gesamtergebnissen, welche auf der x-Achse entsprechend ihrer zugrunde liegenden Lautsprecheranzahl skaliert sind, belegen. Ein großer Anstieg der Qualität zwischen Stereo und 5.1 ist deutlich erkennbar. Hingegen ist der Qualitätsunterschied zwischen den nach MUSHRA *exzellent* bewerteten Zwölf-Kanal-Konfigurationen und der Referenz nur noch gering. Es ist also trotz einer Verdoppelung der Lautsprecheranzahl (12 auf 24) ein nur geringer Anstieg und damit eine Sättigung der Qualität erkennbar. Bezüglich des Vergleichs der Ergebnisse der beiden Wiedergabemethoden lassen sich sowohl an den Graphen, als auch durch einen Wilcoxon-Rangsummentest auf die Differenzbetrachtung keine siginifikanten Unterschiede in der Qualitätsbewertung der Konfigurationen nachweisen.

Die Auswertung der Ergebnisse zeigt, dass auch mit Kunstkopf-BRIRs eine plausible Auralisation möglich ist, in der auch Feinheiten der Unterschiede in der Qualität bewertet werden können. Eine unterschiedliche Bewertung über reale Lautsprecher und Binauralisierung liegt also entgegen der ursprünglichen Annahme nicht vor. Über beide Wiedergabemethoden ist eine deutliche Sättigung der Qualität im Vergleich zur Referenz zu erkennen. Durch die immer noch vorhandene Unterscheidbarkeit zwischen Konfigurationen mit 12 und 24 Kanälen lässt sich keine allgemein gültige Empfehlung zur Einsparung von Rechenleitung und Messaufwand aufstellen. Für viele Anwendungsgebiete könnte jedoch das Übertragen und Berechnen von 12 Kanälen und die daraus resultierende mit exzellent bewertete Qualität ausreichend sein.

Zusammenfassung

In einem Hörversuch wurden die wahrgenommenen Qualitätsunterschiede verschiedener (virtueller) Lautsprecherkonfigurationen untersucht. Die Ergebnisse zeigen keine signifikanten Unterschiede in der Qualitätsabstufung im Vergleich von nicht-individueller Binauralsynthese und realen Lautsprechern. Mit zunehmend größerer Kanalzahl der Konfigurationen zeigt sich ein Sättigungs-Effekt der empfundenen Qualität.

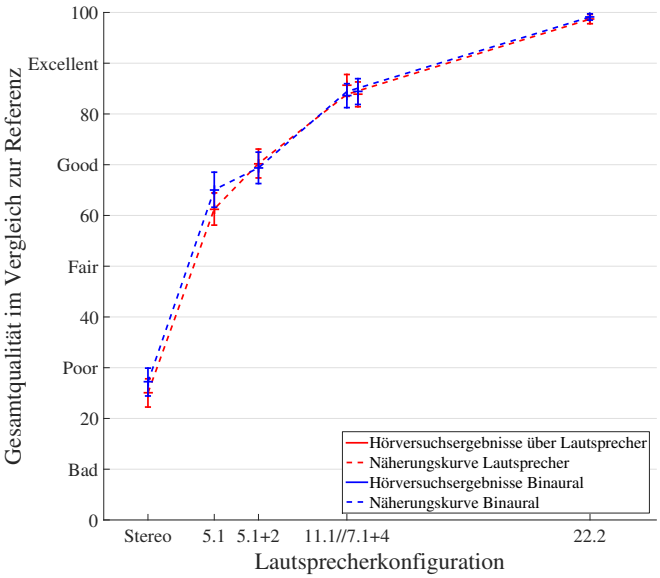

Abbildung 5: Sättigung - Darstellung der Hörversuchs-Gesamtergebnisse skaliert entsprechend der zugrunde liegenden Lautsprecheranzahl

Danksagung

Die Autoren bedanken sich für die hilfreichen Diskussionen und die Zusammenarbeit besonders bei Prof. Dr.-Ing. D. Leckschat und Dr.-Ing. A. Silzle.

Literatur

[1] A. Silzle et al. Vision and Technique Behind the New Studios and Listening Rooms of the Fraunhofer IIS Audio Laboratory. In AES Convention 126, 2009.

[2] W. Hess, J. Weishäupl. Replication of Human Head Movements in 3 Dimensions by a Mechanical Joint. In Proceedings of International Conference on Spatial Audio, ICSA 2014, 2014

[3] M. Geier, J. Ahrens, S. Spors. The SoundScape Renderer: A Unified Spatial Audio Reproduction Framework for Arbitrary Rendering Methods. In Proceedings of 124th AES Convention, 2008.

[4] S. Kim, Y. W. Lee, V. Pulkki. New 10.2-channel Vertical Surround System (10.2-VSS); Comparison study of perceived audio quality in various multichannel sound systems with height loudspeakers. Proceedings of 129th International AES Convention, 2010.

[5] A. Silzle et al. Investigation on the Quality of 3D Sound Reproduction. In Proceedings of International Conference on Spatial Audio, ICSA 2011, 2011.

[6] ISO/IEC 23008-3:2015(E). Information technology – High efficiency coding and media delivery in heterogeneous environments – Part 3: 3D audio. 2015.