

Von mp3 zu PARty:

Wie digitale Signalverarbeitung, Psychoakustik und maschinelles Lernen zusammenfinden

Karlheinz Brandenburg^{1,2}

¹ Institut für Medientechnik, TU Ilmenau, 98693 Ilmenau, E-Mail: karlheinz.brandenburg@tu-ilmenau.de

² Brandenburg-Labs GmbH, 98693 Ilmenau, E-Mail: khb@brandenburg-labs.com

Einleitung

Dieser Beitrag gibt einen Überblick über drei Jahrzehnte Forschung an digitaler Signalverarbeitung in der Audiotechnik. Er kann in der Kürze natürlich nicht vollständig sein, sondern ist geprägt von den persönlichen Erfahrungen des Autors. Ein Thema über all die Zeit ist die Suche nach der „perfekten“ (in Anführungszeichen, da „perfekt“ definiert werden muss) Wiedergabe von Tönen über Lautsprecher oder Kopfhörer.

Die Suche nach perfekter Klangwiedergabe geht zurück bis zur Zeit der ersten technischen Systeme für Tonaufzeichnung und -wiedergabe. Thomas A. Edison organisierte Demonstrationen seiner Technologien, die durch Blindtest (die Beleuchtung des Saales wurde ausgeschaltet, damit die Klangquelle nicht sichtbar war) zeigen sollte, dass die Anwesenden nicht zwischen Original-Tonquelle und der Aufnahme auf einer „Diamond Disc“ seiner Firma unterscheiden konnten (Edison Tone Tests, siehe z.B. [1]).

Geschichte der Audiocodierung

Vor 50 Jahren war „high fidelity“ ein großes Thema in der Tonwiedergabetechnik im Consumer-Bereich. Alles zu diesem Zeitpunkt war analog. Die Digitalisierung des Telefonnetzes (über ISDN, das Integrated Services Digital Network) wurde geplant und dann eingeführt. Um das Jahr 1982 begann die Beschäftigung mit der Idee einer Datenreduktion von Musik. Ausgangspunkt der Arbeiten in Erlangen (es gab unabhängig Arbeiten an verschiedenen Stellen, darunter München (IRT), Duisburg (Uni Duisburg), AT&T Bell Laboratories und andere) war eine Idee von Prof. Dieter Seitzer (Universität Erlangen-Nürnberg), dass doch das digitale Telefonnetz dafür genutzt werden sollte, um Musik in hoher Qualität den Kunden nach Hause zu liefern. Ein Patentprüfer schaute nach dem Stand der Technik und konstatierte (korrekt für die damalige Zeit), dass die 128 kbit/s eines ISDN-Heimanschlusses bei weitem nicht für Musik in hoher Qualität reichen. Prof. Seitzer suchte einen Doktoranden, der schauen sollte, welche Qualität bei welcher Bitrate zu erreichen ist.

In den Jahren 1986/1987 gab es erste Ergebnisse: OCF (Optimum Coding in the Frequency Domain) und andere Verfahren wurden veröffentlicht. Die Grundideen sind seither gleich geblieben:

Das Eingangssignal (z.B. in 44,1 kHz Abtastfrequenz bei 16 bit Auflösung) wird mittels einer Filterbank (je nach

Verfahren 32 bis 2048 Kanäle) in einen Zeit/Frequenzbereich transformiert. Ein sogenanntes Psychoakustisches Modell schätzt für jeden Filterkanal für jeden Datenblock ab, wie groß der „masked threshold“, also die zeit- und frequenzaktuelle Maskierungsschwelle ist. Daraus kann z.B. eine erlaubte Quantisierungsstufengröße errechnet werden. Es folgt die Quantisierung und Codierung, die sicherstellen soll, dass der Quantisierungsfehler nicht die Schwelle der Hörbarkeit überschreitet. Die nachfolgende Abbildung ist aus dem MPEG-1-Standard entnommen und zeigt die grundsätzliche Idee, der immer noch alle (auch nachfolgenden) Audiocodierverfahren folgen.

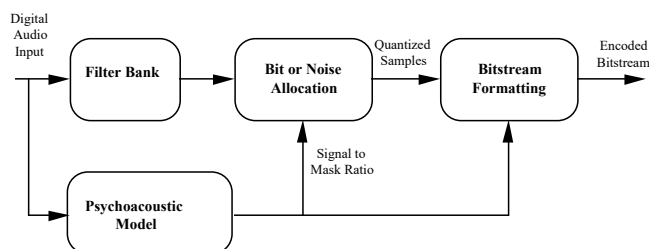


Abb. 1: Allgemeines Schema einer gehörangepassten Audiocodierung (Quelle: ISO/IEC IS 11172-3 [2])

Ein ganz wesentlicher Schritt hier ist die Abschätzung der Maskierungsschwellen. Der Autor folgte damals dem in den 80er Jahren frisch erschienenen Buch von Prof. Zwicker [3]. Ein wesentlicher Bestandteil dieser Modellierung ist die sogenannte Maskierung im Frequenzbereich. Ein (hier nachgezeichnetes) Bild aus dem Buch von Prof. Zwicker war damals in jedem Zeitschriften- und Konferenzbeitrag zur Audiocodierung zu finden.

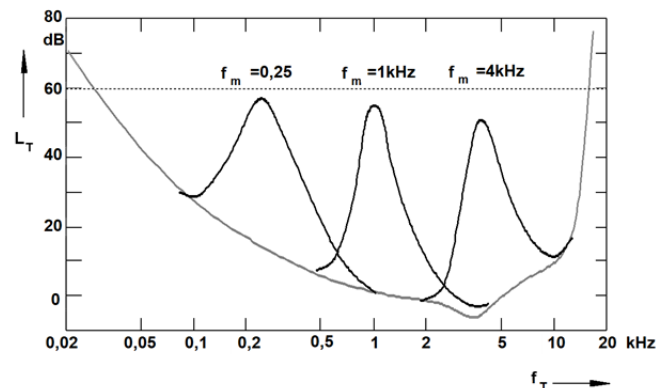


Abb. 2: Maskierung im Frequenzbereich (nachgezeichnet aus [3])

Diese Vorgehensweise war damals in gewissem Sinne revolutionär: In der digitalen Signalverarbeitung und Nachrichtentechnik wurden komplexe Systeme als linear betrachtet und entsprechend im euklidischen Raum optimiert. Als Fehlermaß wurde also ein quadratischer Abstand verwendet, in der Audiotechnik als SNR, Signal-Rausch-Abstand bekannt. Die Optimierung bezüglich der Hörbarkeit von Quantisierungsfehlern ist deutlich unterschiedlich von einem quadratischen Fehlermaß. Ein Korollar daraus ist, dass gut klingende Audiocodiersysteme einen schlechteren SNR verglichen mit der klassischen Optimierungsstrategie zeigen. Dies führte dazu, dass es keine Messverfahren gab und jede Verbesserung eines Audiocodierverfahrens per aufwändigem Hörtest verifiziert werden musste.

Die Entwicklung der Audiocodierung bekam einen massiven Schub durch zwei Ereignisse: Auf Vorschlag des Instituts für Rundfunktechnik in München gab es ein pan-europäisches Projekt (DAB, Digital Audio Broadcasting, Eureka 147), in dem die Entwicklung einer Audiocodierung bei niedrigen Bitraten eine Schlüsselstellung einnahm. Kurze Zeit später startete die weltweite Standardisierung von Audiocodierverfahren im Rahmen der MPEG (Moving Pictures Experts Group), die zunächst die Codierung von Bewegtbildern bei einer Bitrate bis zu 1,5 Mbit/s als Ziel hatte, allerdings auch schon die Aufzeichnung von Musik auf Festplatten und die Übertragung über Digitalradio oder Computernetzwerke in ihrer Liste der Anwendungsmöglichkeiten nannte. Das Ergebnis dieser Standardisierung war ein Audiocodiersystem (MPEG Audio) mit drei Modi (Layer I, Layer II und Layer III). Von diesen kamen Layer II und Layer III (später nach der für die Speicherung auf Computern benutzten Fileendung mp3 genannt) zu großer Verbreitung. Selbst heute können praktisch alle Geräte, die Musik spielen (einschließlich sämtlicher Mobiltelefone) Layer II und Layer III wiedergeben.

MPEG-1 Audio wurde 1992 als Standard festgeschrieben (siehe [4]). MPEG-2 Audio bekam zunächst nur minimale Erweiterungen zur MPEG-1-Syntax („backwards compatible“) und später, als System der zweiten Generation, mit AAC (Advanced Audio Coding) einen auf denselben Grundideen aufbauenden würdigen Nachfolger (bessere Qualität, niedrigere Bitraten bei gleicher Qualität möglich). Die Beschreibung der weiteren Verfahren ist hier nicht möglich, dafür sei (für eine Auswahl) auf [5] verwiesen.

Kognitive Effekte

Schon früh war den Verfechtern der gehörangepassten Audiocodierung klar, dass nicht alle hörbaren Effekte durch die Mechanik des Ohres begründet werden können. Insbesondere bei der Perzeption von räumlichem Schall spielen höhere Schichten in unserem Gehirn eine wesentliche Rolle. Die folgende Abbildung (private Kommunikation von James D. Johnston) gibt einen Überblick über die Verarbeitung einschließlich der Rückkopplung bis hin zum Ohr.

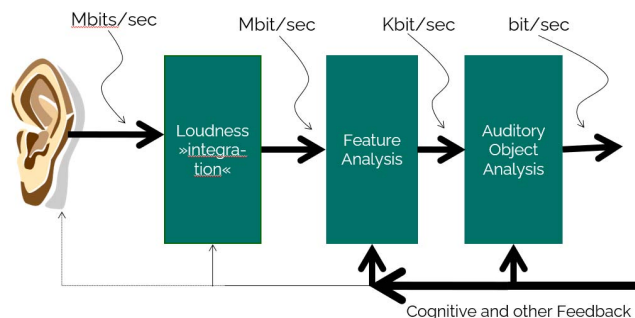


Abb. 3: Audioverarbeitung in Ohr und Gehirn (Bildquelle: James D. Johnston)

Bekannte kognitive Effekte sind z.B. der Bauchrednereffekt (Lokalisierung des Gehörten auf die Puppe, obwohl sie natürlich nicht spricht) oder der McGurk-Effekt (bei Divergenz zwischen visuellem und Hör- Eindruck werden andere Silben erkannt). Grundsätzlich gilt, dass unser Gehirn über ein sehr gutes „pattern matching“ verfügt und deshalb jeder Höreindruck auf andere, gelernte Eindrücke bezogen werden kann. Wie das im Detail geschieht, ist auch nach Jahrzehnten der Forschung noch umstritten.

Fast forward: Was sind aktuelle Themen in der Audiosignalverarbeitung?

An dieser Stelle soll nur auf eine kleine Auswahl an Themen eingegangen werden:

- Music Information Retrieval (MIR)
- Räumliche Klangwiedergabe über Lautsprecher
- Räumliche Klangwiedergabe über Kopfhörer
- PARty (Personalized Auditory Reality)

Music Information Retrieval (MIR)

Seit mehr als 20 Jahren gibt es Forschung zur automatischen Analyse von Musik. Die Ergebnisse werden sowohl in den Musikwissenschaften angewandt als auch in aktuellen Streaming-Systemen verwendet, um Hörer:innen angepasste Empfehlungen zu geben. „Spiel mir meine Lieblingsmusik“ oder „Stelle mir eine Playlist ähnlich der Musik dieser Interpretin zusammen“ sind konkrete Fragestellungen in diesem Bereich. Eine Variante ist „Query by Humming“, also die Erkennung von Musik nach einem Vorsingen/Vorsummen einer Melodie. Abbildung 4 zeigt einen Prototyp eines solchen Gerätes. Die weitverbreitete Anwendung dieser Technologien scheiterte aber daran, dass nur wenige Personen ungefähr erinnerte Melodien gut genug singen oder summen können. Eine Übersicht über das Gebiet des MIR gibt es z.B. im Buch von Meinard Müller [6].



Bild 4: Prototyp einer Musikererkennungshardware für den Einsatz in einem Musikgeschäft (Bildquelle: Fraunhofer IDMT)

Räumliche Klangwiedergabe über Lautsprecher

Um Klang im Raum mit hoher Plausibilität wiedergeben zu können, reicht klassische Surround-Technik nicht. In der Zwei-Kanal-Stereotechnik und auch in klassischen Surroundverfahren wird genutzt, dass Menschen eine grobe Ortung eines Schalls aufgrund der Intensitätsunterschiede z.B. zwischen rechtem und linkem Lautsprecher wahrnehmen können. Eine bessere Audio-Illusion erfordert unter anderem, dass die Phasen richtig wiedergegeben werden. Dazu gibt es verschiedene Ansätze, zu denen „Vector Based Amplitude Panning“ (VBAP), die Wellenfeldsynthese und Ambisonics gehören. Den letzten beiden Varianten ist gemeinsam, dass versucht wird, mittels mehreren bis vielen Lautsprechern die Wellenfronten im Zuhörerbereich korrekt zu synthetisieren. Der Ausgangspunkt dafür ist eine Ableitung der Wellengleichung mittels des Kirchhoff-Helmholtzintegrals. Mit Hilfe dieser Integral-Identität können die Druck- und Schnelle-Daten an einem beliebigen Punkt im Zuhörerraum aus den Daten auf einer Oberfläche, die den Zuhörerraum umschließt, hergeleitet und damit auch synthetisiert werden. Als Beispiel soll hier nur Ambisonics vorgestellt werden: Die Grundidee ist die Darstellung des Schallfelds durch spherical harmonics. Bild 5 erklärt das Prinzip. Das Schallfeld wird aus räumlichen Komponenten rekonstruiert:

Zunächst der ungerichtete Schalldruck, dann Amplituden in $x/y/z$ – Richtung und dann in entsprechend höheren Ordnungen. Als „B-Format“ wurde zunächst nur eine Darstellung erster Ordnung eingesetzt. Dazu gibt es auch Mikrofon-Anordnungen, die direkt ein solches Signal aufnehmen. Heute wird an vielen Stellen „higher order ambisonics“ eingesetzt. Erst mit höheren Ordnungen kann eine plausible Wiedergabe außerhalb eines sogenannten Sweet Spot ermöglicht werden. Ambisonics findet z.B. als eine Option der internen Signaldarstellung Anwendung im MPEG-H-Standard (siehe z.B. [7]). Ein ausführlicher Überblick ist in [15] zu finden.

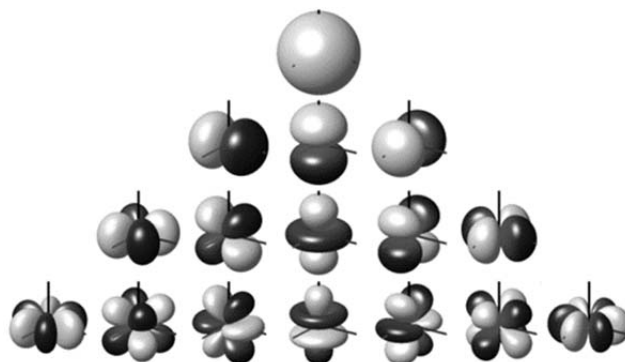


Bild 5: Prinzipdarstellung Ambisonics: Von oben nach unten Ordnung der sphärischen Harmonischen, von links nach rechts die verschiedenen Raumrichtungen in einer Ordnung. (Bildquelle: Franz Zotter, IEM Graz)

Räumliche Klangwiedergabe über Kopfhörer

Während aktuell hervorragende Klangillusionen per Lautsprecher möglich sind, gibt es für das Äquivalent über Kopfhörer seit Jahrzehnten Fortschritte, die aber nicht wirklich zufrieden stellen. In der Grundversion (Zwei-Kanal-Stereophonie direkt über Kopfhörer wiedergegeben) scheinen alle Klangquellen im eigenen Kopf positioniert zu sein. Verbesserte Binaural-Technologien (siehe unten) zeigen weiterhin eine verminderte Plausibilität, z.B. findet sich häufig ein Problem mit Vorne-Hinten-Vertauschung. Die Wiedergabe über Lautsprecher von z.B. einer traditionellen Stereoanlage wird als natürlicher empfunden.

Die erste, wirklich beeindruckende Technologie zur realitätskonformen Wiedergabe von Klang über Kopfhörer war die Kunstkopfstereophonie. Zur Aufnahme wurden in einen Kunstkopf (z.B. Head-and-Torso-Simulator, HATS), an der Stelle der Trommelfelle, Mikrophone eingebaut. Deren Signal enthält also alle Phasen- und Amplitudeninformationen wie im Referenz-Raum. Selbst eine (nicht an den Hörer angepasste) Außenohrüber-tragungsfunktion (HRTF, Head Related Transfer Function) ist automatisch berücksichtigt.

Als nächster Schritt wurden HRTFs des Hörers / der Hörerin gemessen und zum binauralen Rendering einbezogen.

Ein deutlicher Schritt zu besserer räumlicher Illusion war die Einbeziehung von Head Tracking (erste bekannte Demo war das Convolvotron Ende der 80er Jahre, siehe [8]). Damit wird auch bei Kopfdrehungen die jeweilig richtige HRTF

eingesetzt, das Ergebnis ist deutlich natürlicher als ohne Verwendung von Head Tracking.

Als weitere Variante wurde noch eine Raumsimulation auf das Signal gerechnet, also nicht nur eine HRTF, sondern eine BRIR (Binaural Room Impulse Response) verwendet.

Weitere Parameter wurden einbezogen, darunter auch Informationen über den tatsächlichen Wiedergaberaum (siehe unten). Nach heutigem Stand gibt es eine Reihe von Elementen, die eine plausible Wiedergabe über Kopfhörer besser ermöglichen. Es ist noch nicht klar, wie diese Parameter zu gewichten sind. Dazu gehören

- Die persönlichen Anatomie (HRTF etc.)
- Raum-Cues (Einbeziehung von Reflexionen)
- Visuelle Cues
- Proprioception, also das Gefühl für den eigenen Körper im Raum
- Eigener Erfahrung, Wissen über die Eigenschaften der Klangquellen und der Hörumgebung.

Als Ziel solcher Wiedergabe können Authentizität (Übereinstimmung mit einer externen Referenz, siehe [9]) oder Plausibilität, also die Übereinstimmung mit einer gelernten Referenz (real oder in besonderen Fällen simuliert) gefragt sein. Die inneren Referenzen sind dabei nicht unproblematisch, sie können individuell unterschiedlich, ungenau oder gar nicht vorhanden sein (siehe [10]).

In verschiedenen Versuchen wurde der sogenannte „room divergence effect“ klar experimentell verifiziert: Wenn in einem Raum Aufnahmen über Kopfhörer wiedergegeben werden, die mit BRIRs entweder aus demselben Raum oder einem anderen Raum beaufschlagt sind, dann ergibt sich eine massive Verbesserung der Externalisierung bei kongruenter Wiedergabe, insbesondere in weniger trockenen Räumen. Externalisierung meint hier den Effekt, dass eine virtuelle Schallquelle deutlich entfernt vom eigenen Kopf wahrgenommen wird. Dieser Effekt war vom Hörensagen her seit Jahrzehnten bekannt und wurde aber erst in der Dissertation von Stephan Werner [11] in mehr Detail erforscht.

Ein weiterer oft vernachlässigter Effekt bei dem Versuch, räumliches Audio über Kopfhörer wiederzugeben, ist die Adaptation des Höreindrucks über die Zeit hinweg. Diese Effekte wurden in der Dissertation von Florian Klein [12] in etlichen Experimenten untersucht. Bild 6 zeigt den Effekt: Wenn die Hörer auf einen Raum eingehört sind, verbessert sich dort der Eindruck der Externalisierung.

Die Gruppe SR hat zunächst den schallharten Raum SR gehört, Gruppe LL den trockenen Raum LL. Die virtuelle Wiedergabe des Raums LL wird deutlich unterschiedlich beurteilt je nach Trainingsphase.

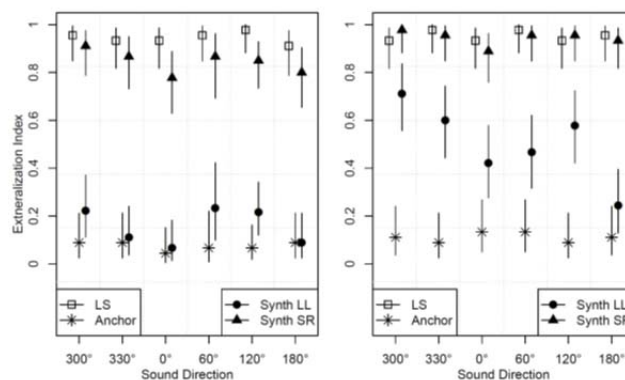


Bild 6: Versuch zu Lern- und Trainingseffekten. Y-Achse mittlerer Externalisierungsindex. Aus F. Klein und S. Werner [13]

Weitere Forschungen in diesem Bereich betreffen die Plausibilität der Wiedergabe bei Einbeziehung eigener Bewegung (besser, siehe [14]) sowie die Frage, wie die BRIRs gemessen werden können. Eine Messung in genügender Auflösung (z.B. alle 25 cm) ist für praktische Anwendungen aus Aufwandsgründen nicht geeignet. Es wurden Interpolationsverfahren entwickelt, mit denen eine oder einige wenige Ausgangspositionen ausreichend sind. Weitere Informationen finden sich z.B. in [16], neuere Überblicksveröffentlichungen unter Beteiligung des Autors in [17] und [18].

Die Zukunft: PARTy (Personalized Auditory Reality)

Wenn einmal Technologien für eine wirklich plausible Wiedergabe durch binaurales Rendering zur Verfügung stehen, können diese mit anderen Technologien (Auditory Scene Analysis, Source Separation und Realtime Analysis of the Room) zu zukünftigen intelligenten Kopfhörersystemen verbunden werden. Bild 7 zeigt diese Vision:

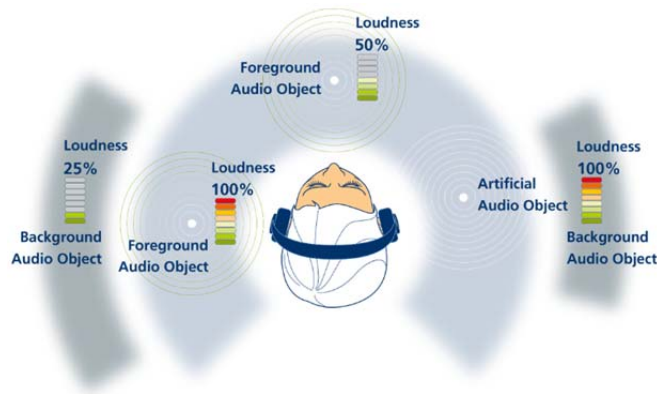


Bild 7: Funktionen eine PARTy-Systems (Bildquelle: TU Ilmenau)

Im Beispiel einer Cocktail-Party-Situation soll dieses System störenden Schall (von anderen Gästen) verringern und den Schall der Personen, mit denen ich reden will, verstärken. Ebenso kann ein Telefongespräch eingespielt werden, als sei die andere Person im Raum in meiner Nähe. Das geschieht in einer akustisch transparenten Weise, d.h. ich habe nicht wirklich das Gefühl, mit einem Kopfhörer herumzulaufen.

Um zu solchen Systemen zu gelangen, sind noch viele Herausforderungen zu lösen (siehe z.B. [19]). Nicht die geringste Schwierigkeit dabei sind die Echtzeitanforderungen, gerade für den Teil der Verringerung von Störschall. Benötigt werden Fortschritte in der Erkennung von Schall mit Hilfe von KI, denn es sind theoretisch unendlich viele Klassen von Klängen zu unterscheiden, aber alles soll in Millisekunden gerechnet werden. Auch die bisherigen Beispiele von Source Separation erfüllen nicht die Anforderungen an die Tonqualität eines solchen Systems.

Zusammenfassung und Ausblick

Digitale Signalverarbeitung und das Wissen über die Funktion von Gehör und Gehirn haben über die letzten Jahrzehnte wesentliche Fortschritte gebracht. Wenn wir heute Musik hören, dann sind so entstandene Algorithmen wesentlich beteiligt. Angesichts aktueller Arbeiten an vielen Stellen in der Welt kann konstatiert werden: Wir sind noch lange nicht fertig mit diesen Technologien, die Wissenschaft und Technik der Ton-Aufzeichnung und – Wiedergabe wird noch wesentlich weiterentwickelt werden.

Acknowledgements:

Die hier beschriebenen Ilmenauer Arbeiten wurden teilfinanziert von der Deutschen Forschungsgemeinschaft (BR 1333/13-1 und BR 1333/18-1), dem Freistaat Thüringen sowie dem ESF (European Social Fund).

Literatur

- [1] Library of Congress, URL: <https://blogs.loc.gov/now-see-hear/2015/05/is-it-live-or-is-it-edison/>
- [2] ISO/IEC 11172-3: 1993 Information technology - Coding of moving pictures and associated audio for digital storage media at up to about 1,5 Mbit/s - Part 3: Audio. ISO-SC 29
- [3] Zwicker, E.: Psychoakustik. Springer-Verlag, Berlin, 1982
- [4] Brandenburg, K., Stoll, G. et. al: The ISO/MPEG-Audio Codec: A Generic Standard for Coding of High Quality Digital Audio. J. of the Audio Eng. Soc (1994), 780 - 792
- [5] Brandenburg, K., Faller, C., Herre, J., Johnston, J. D., & Kleijn, W. B. (2013). Perceptual coding of high-quality digital audio. Proceedings of the IEEE, 101(9), 1905-1919.
- [6] Müller, Meinard. Fundamentals of music processing: Audio, analysis, algorithms, applications. Springer, 2015.
- [7] Herre, J., Hilpert, J., Kuntz, A., & Plogsties, J. (2015). MPEG-H audio—the new standard for universal spatial/3D audio coding. Journal of the Audio Engineering Society, 62(12), 821-830.
- [8] Wenzel, E. M., Wightman, F. L., Kistler, D. J., & Foster, S. H. The convolotron: Realtime Synthesis of out-of-head localization, 1988. In Joint Meeting of the Acoustical Society of America and the Acoustical Society of Japan, Honolulu, November (pp. 14-18).
- [9] Brinkmann, F. et al.: A round robin on room acoustical simulation and auralization. J. Acoust. Soc. Am. 145 (2019), 2746
- [10] Neidhardt, A., Zerlik, A. M.: The Availability of a Hidden Real Reference Affects the Plausibility of Position-Dynamic Auditory AR. Frontiers in Virtual Reality 2 (2021), 102
- [11] Werner, S.: Über den Einfluss kontextabhängiger Qualitätsparameter auf die Wahrnehmung von Externalität und Hörereignisort. Dissertation, TU Ilmenau, 2019
- [12] Klein, F.: Auditive Adaptationsprozesse im Kontext räumlicher Audiowiedergabesysteme. Dissertation, TU Ilmenau, 2021
- [13] Klein, F., Werner, S.: Influences of training on externalization in binaural synthesis in situations of room divergence. Journal of the Audio Engineering Society 65/3 (2017)
- [14] Neidhardt, A., et al.: Plausibility of an interactive approaching motion towards a virtual sound source. AES Conv. Milan, Italy, 2018
- [15] Zotter, F.; Frank, M. Ambisonics—A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality, 1st ed.; Springer: Heidelberg, Germany, 2019.
- [16] Sloma, U.; Klein, F.; Werner, S.; Pappachan Kannookadan, T.: Synthesis of Binaural Room Impulse Responses for Different Listening Positions Considering the Source Directivity. In Proceedings of the 147th International AES Convention, New York, NY, USA, 16–19 October 2019.
- [17] Brandenburg, K., Klein, F., Neidhardt, A., Sloma, U., & Werner, S. (2020).: Creating auditory illusions with binaural technology. In The Technology of Binaural Understanding (pp. 623-663). Springer, Cham
- [18] Werner S, Klein F, Neidhardt A, Sloma U, Schneiderwind C, Brandenburg K.: “Creation of Auditory Augmented Reality Using a Position-Dynamic Binaural Synthesis System—Technical Components, Psycho-acoustic Needs, and Perceptual Evaluation” MDPI Applied Sciences. 2021; 11(3):1150. <https://doi.org/10.3390/app11031150>
- [19] Brandenburg, K.; Cano, E.; Klein, F.; Köllmer, T.; Lukashovich, H.; Neidhardt, A.; Sloma, U.; Werner, S. Plausible Augmentation of Auditory Scenes using Dynamic Binaural Synthesis for Personalized Auditory Realities. In Proceedings of the Audio Engineering Society Conference: 2018 AES International Conference on Audio for Virtual and Augmented Reality, Redmond, WA, USA, 20–22 August 2018.