Perzeptive Untersuchung zur Mixing Time und deren Einfluss auf die Auralisation

Philipp Stade[1,2]
[1] FH Köln, Institut für Nachrichtentechnik
[2] Technische Universität Berlin, Fachgebiet Audiokommunikation
E-Mail: philipp.stade@fh-koeln.de

Einleitung

Die *Mixing Time* kennzeichnet den Zeitpunkt, ab dem das Schallfeld in einem Raum in einen physikalisch diffusen Nachhall übergeht. Dementsprechend wird als *perzeptive Mixing Time* der Zeitpunkt bezeichnet, ab dem dieses Schallfeld diffus wahrgenommen wird und keine Unterscheidung zwischen verschiedenen Orientierungen mehr möglich ist. Sie ist unter anderem von hoher Bedeutung für die dynamische Binauralsynthese. Ein voll diffuser Nachhall erfordert in der Auralisation keine Anpassung des Schallfelds an die Kopfdrehung. Somit kann die Kenntnis der perzeptiven Mixing Time für die Daten- und Rechenaufwandsreduktion in einem dynamischen Binauralsystem genutzt werden.
Die perzeptive Mixing Time war bereits Gegenstand zahlreicher Untersuchungen, z.B. [1][2]. Letztere beschränken sich auf Schwellwertuntersuchungen zur Bestimmung eines absoluten Zeitpunktes. Hierbei ist die Ununterscheidbarkeit zum Referenzstimulus das Abbruch-Kriterium der Experimente. Eigene Voruntersuchungen mit ähnlicher Vorgehensweise führten jedoch zu abweichenden Ergebnissen. Mitunter konnten noch bei sehr späten Überblendzeitpunkten Unterschiede erkannt werden, sodass sich keine Ununterscheidbarkeit einstellte. Daher nimmt sich die vorliegende Untersuchung zum Ziel, verschiedene Überblendzeiten in einen statischen Nachhall eher qualitativ zu bewerten, um deren Einfluss auf die Auralisation zu untersuchen. Häufig ist in praktischen Anwendungen keine Ununterscheidbarkeit notwendig, sondern eine ausreichend hohe Ähnlichkeit hinreichend.
In einem Hörversuch mit dynamischer Binauralsynthese erfolgt ein Vergleich von voll-dynamischen binauralen Raumimpulsantworten (BRIRs) und BRIRs mit statischer Nachallfahne. Dabei werden verschiedene Übergangszeitpunkte herangezogen und beurteilt. Es wird weiterhin überprüft, inwieweit der Direktschall als Maskierer wirkt und die Bewertung beeinflusst. Dazu werden zu den jeweiligen Zeitpunkten die dynamischen und statischen Nachhallfahnen gegenübergestellt und bewertet. Zudem wird untersucht, ob die Richtcharakteristik der Anregungsquelle eine Auswirkung auf die Bewertung hat, daher werden omnidirektionale und gerichtete Schallquellen miteinander verglichen und beurteilt.

Räume

Für die perzeptive Untersuchung wurden zwei Konzertsäle des WDR Funkhauses in Köln (*kleiner Sendesaal* und *Klaus-von-Bismarck-Saal*) mit unterschiedlicher Größe und Nachhallzeit ausgewählt (siehe Tabelle 1). In beiden Räumlichkeiten wurden mit einem Kunstkopf (*Neumann KU100*) gedrehte binaurale Raumimpulsantworten in 1°-Auflösung erfasst. Weiterhin wurde die Anregungsquelle variiert und bei gleicher Empfängerposition wurden die Impulsantworten jeweils mit omnidirektionaler Schallquelle und gerichtetem Lautsprecher gemessen [6].
In der Literatur finden sich model- (Analyse der Geometrie) als auch datenbasierte (Analyse der Impulsantwort) Prädiktoren zur Bestimmung der perzeptiven Mixing Time. Basierend auf ausgewählten Untersuchungen liegen diese Zeitpunkte beim kleinen Sendesaal zwischen 35 ms und 74 ms und beim Klaus-von-Bismarck-Saal zwischen 74 ms und 140 ms, siehe Tabelle 2.

Raum	**V**	**A**	**S**	**RT**	α
kl. Sendesaal	1247 m^3	204 m^2	750 m^2	0.83 s	0.32
KVB-Saal	6098 m^3	480 m^2	2777 m^2	1.46 s	0.24

Tabelle 1: Volumen V, Grundfläche A, Oberfläche S, mittlere Nachhallzeit RT und mittlerer Absorptionskoeffizient α der auralisierten Räume.

Raum	**Cr**	**Ru**	**Hi**	**Li**$_{model}$	**Li**$_{data}$
kl. Sendesaal	35	77	74	65	62
KVB-Saal	78	102	140	121	74

Tabelle 2: perzeptive Mixing Times in ms, basierend auf ausgewählter Literatur (**Cr**emer [3], **Ru**bak [4], **Hi**daka [5], **Li**ndau [1]).

Versuchsumgebung

Die Hörversuche wurden kopfhörerbasiert (*AKG K601*) in einer virtuellen auditiven Umgebung mit Hilfe eines dynamischen Binauralrenderers (*SoundScape Renderer* [7]) durchgeführt. Die Kopfposition der Probanden wurde auf der Horizontalebene in 1°-Auflösung erfasst (Headtracker: *Polhemus FastTrack*) und im Binauralrenderer eine Faltung der nachhallfreien Signale mit den entsprechenden BRIR-Datensätzen in Echtzeit durchgeführt. Die Verwendung eines Drehstuhls ermöglichte den Probanden während des Versuchs eine freie Drehung in der Horizontalebene. In einem Vorgespräch wurden die Probanden ausdrücklich dazu angehalten, die Orientierung mehrfach zu wechseln. Um Beeinträchtigungen durch Störgeräusche zu vermeiden, fanden die Versuche im reflexionsarmen Raum der FH Köln statt, der Abhörpegel betrug 75 dB(A) SPL. Eine nachhallfreie Schlagzeug-Aufnahme bestehend aus Kick, Snare und Hi-Hat wurde als Testsignal verwendet. Die allge-

meine Vorgehensweise des Versuchsaufbaus hat sich bereits in vorherigen Untersuchungen bewährt [8]. Die gerenderten Stimuli wurden kontinuierlich wiederholt und die verschiedenen BRIR-Sätze konnten unmittelbar umgeschaltet werden. Das Versuchs-Design wurde in Anlehung an einen „MUlti Stimulus test with Hidden Reference and Anchor“ (MUSHRA) durchgeführt [9]. Dabei sollten die Probanden die globale Ähnlichkeit der dargebotenen Stimuli zur Referenz auf der kontinuierlichen Qualitäts-Skala (*continuous quality scale - CQS)* von 0 - 100 und den entsprechenden Attributen (*bad*, *poor*, *fair*, *good* und *excellent*) bewerten. Dabei entspricht eine Bewertung von 100 einer Ununterscheidbarkeit zum Referenzstimulus. Abweichend von der ITU-Empfehlung wurde als Anker eine Mono-Impulsantwort aus den jeweiligen Räumen verwendet. Als Hörversuchssoftware kam *Scale* zum Einsatz, die eine direkte Anbindung an den *SoundScape Renderer* bietet [10]. Die Eingabe der Bewertungen erfolgte mit Hilfe eines Tablet-PCs. Die nachfolgend erläuterten BRIR-Datensätze wurden randomisiert in einem gemeinsamen Hörversuch auralisiert und bewertet. Die Datensätze von Experiment 1 (LEA) konnten pro Raum und Anregungsquelle jeweils auf einem gemeinsamen Fragebogen bewertet werden, die Datensätze von Experiment 2 (LEB) mussten aufgrund der veränderten Referenz (abhängig von der Übergangszeit) jeweils separat untersucht werden. Insgesamt nahmen 16 Probanden mit einem Durchschnittsalter von 30 Jahren an zwei Terminen teil, 9 Probanden hatten bereits Erfahrungen mit ähnlichen Hörversuchen.

Raum	**1**	**2**	**3**	**4**	**5**	**6**
kl. Sendesaal	5	10	20	40	80	160
KVB-Saal	10	20	40	80	160	320

Tabelle 3: Sechs untersuchte Überblendzeiten T (in ms) in einen statischen Nachhall

BRIR Datensätze

Die für die Auralisierung mittels dynamischer Binauralsynthese verwendeten BRIR-Datensätze basieren auf den Impulsantworten der zuvor beschriebenen Kunstkopf-Messungen in den entsprechenden Räumlichkeiten. Diese Impulsantworten wurden ab verschiedenen Übergangszeiten in einen statischen Nachhall übergeblendet. Ingesamt wurden pro Hörversuch, Raum und Anregungsquelle sechs verschiedene Zeitpunkte verwendet, sodass 48 modifizierte BRIR-Datensätze zu bewerten waren. Bezugnehmend auf Prädiktoren der Mixing Time vorheriger Untersuchungen (siehe Tabelle 2) wurden die Zeitpunkte aus Tabelle 3 ausgewählt. Dabei galt es den Bereich zu frühen und späten Zeitpunkten zu erweitern. Eine Reduzierung auf sechs einzelne Überblenzeitpunkte war erforderlich, um zu umfangreiche MUSHRA Fragebögen zu vermeiden.

Experiment 1 (LEA)
Als Referenz wurde in diesem Hörversuch der originale, gemessene BRIR-Datensatz verwendet (siehe Abbildung 1a), der eine voll-dynamische Anpassung an die Kopfdrehung erlaubt. Bei den modifizierten Vergleichs-Datensätzen wurde ab den Zeitpunkten aus Tabelle 3 die dynamische Anpassung unterbunden, indem für den gesamten Winkelbereich ausschließlich der entsprechende Zeitbereich der frontal orientierten BRIR (0°-Position) verwendet wurde (siehe Abbildung 1b). Desweiteren wurde eine Energieanpassung der Datensätze untereinander durchgeführt, um einen Einfluss der Lautstärke auf die Bewertung auszuschließen.

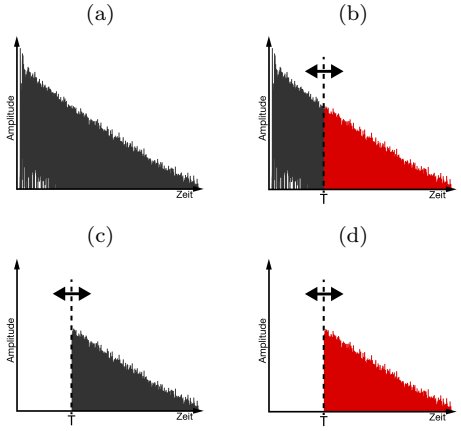

Abbildung 1: Referenz LEA (a) und LEB (c), Modifikation LEA (b) und LEB (d); schwarz = dynamisch, rot = statisch; ab T keine Anpassung an die Kopfdrehung, T variabel.

Experiment 2 (LEB)
Als Basis für die Referenz wurde in diesem Hörversuch ebenfalls der originale und messtechnisch erfasste, dynamische BRIR-Datensatz verwendet. Alle Anteile vor den Zeitpunkten aus Tabelle 3 wurden jedoch aus den Impulsantworten entfernt, sodass nur die Nachhallfahnen auralisiert wurden (siehe Abbildung 1c). Die Vergleichs-Datensätze wurden gleichermaßen gekürzt, analog zu LEA wurde hier als statische Nachhallfahne ausschließlich die Nachhallfahne der frontal orientierten BRIR (0°-Position) verwendet (siehe Abbildung 1d). Eine weitere Pegel-Anpassung dieser Datensätze (zusätzlich zur Energieanpassung aus LEA) wurde nicht durchgeführt, um die beiden Experimente direkt miteinander vergleichen zu können und eventuell auftretende Maskierungseffekte nicht zu beeinflussen.

Auswertung

In der folgenden Auswertung der Experimente werden die Mittelwerte der Bewertungen betrachtet. Die statistische Signifikanz der Bewertungen wurde mittels eines Zweistichproben-t-Test mit einem Signifikanzniveau von $\alpha = 0.05$ bzw. $\alpha = 0.01$ überprüft. Hierbei wurde das Alphaniveau aus Gründen der Alphafehler-Kumulierung mit Hilfe der Hochberg-Prozedur [11] korrigiert. Dabei wurde zum einen überprüft, ob die Bewertungen der einzelnen Zeitpunkte signifikant unterschiedlich zu den Bewertungen der Referenz sind und zum anderen, ob bei gleichen Zeiten die Bewertungen zwischen den Richtcharakteristika sowie den Versuchen mit (LEA) und ohne (LEB) Direktanteil signifikant unterschiedlich zueinander sind.

Raum	1	2	3	4	5	6
kl. Sendesaal	0.02	0.006	0.03	0.8	0.8	0.3
KVB-Saal	0.03	0.03	0.6	0.7	0.1	0.9

Tabelle 4: Vergleich der Bewertungen omnidirektionale vs. gerichtete Quelle; Signifikanz-Werte aus Zweistichproben-t-Test je Übergangszeitpunkt 1-6

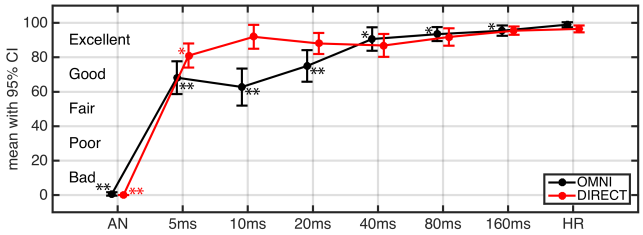

Abbildung 2: Bewertete Ähnlichkeit LEA, kleiner Sendesaal, omnidirektionale (schwarz) vs. gerichtete (rot) Quelle; signifikant unterschiedlich zur Referenz: * ($\alpha = 0.05$) bzw. ** ($\alpha = 0.01$)

Die nachfolgenden Grafiken (Abbildungen 2-5) zeigen die Mittelwerte der Bewertungen mit 95 % Konfidenzintervallen, AN bezeichnet hierbei den Anker (Mono-Impulsantwort aus den jeweiligen Räumen ohne Headtracking), HR die versteckte Referenz. Auf der X-Achse sind zudem die sechs verwendeten Übergangszeitpunkte in den statischen Nachhall aufgetragen, die Y-Achse zeigt die MUSHRA Bewertungs-Skala mit ihren entsprechenden Attributen.

Experiment 1 (LEA)

In beiden Räumlichkeiten zeigt sich bei Verwendung der omnidirektionalen Schallquelle tendenziell eine mit ansteigender Übergangszeit in den statischen Nachhall ansteigende Bewertung der Stimuli (siehe Abbildungen 2 und 3, schwarz). Alle untersuchten Zeitpunkte bis 160 ms wurden signifikant schlechter im Vergleich zur Referenz bewertet. Im Klaus-von-Bismarck-Saal zeigt sich bei Verwendung der gerichteten Quelle ein ähnlicher Verlauf (siehe Abbildung 3, rot), auch hier wurden alle Stimuli bis 160 ms signifikant schlechter im Vergleich zur Referenz bewertet. Im kleinen Sendesaal hingegen wurde bei Verwendung der gerichteten Quelle nur der Anker sowie der Zeitpunkt von 5 ms als signifikant unterschiedlich im Vergleich zur Referenz bewertet (siehe Abbildung 2, rot). Stellt man die Bewertungen für die Quellen bei gleichen Zeitpunkten gegenüber und vergleicht die Richtcharakteristika, so zeigen sich im kleinen Sendesaal für die Zeitpunkte 5 ms, 10 ms und 20 ms bzw. im Klaus-von-Bismarck-Saal für die Zeitpunkte 10 ms und 20 ms signifikante Unterschiede (siehe Tabelle 4). Dabei wurde die omnidirektionale Quelle jeweils signifikant schlechter als die gerichtete Quelle bewertet.
Die Mittelwerte der Bewertungen liegen im kleinen Sendesaal für beide Anregungsquellen bereits ab einem Zeitpunkt von 5 ms bei CQS > 60 und entsprechen somit dem Attribut *good*. Im Klaus-von-Bismarck-Saal wird diese Schwelle bei 40 ms (für die omnidirektionale Quelle) bzw. bei 10 ms (für die gerichtete Quelle) überschritten.

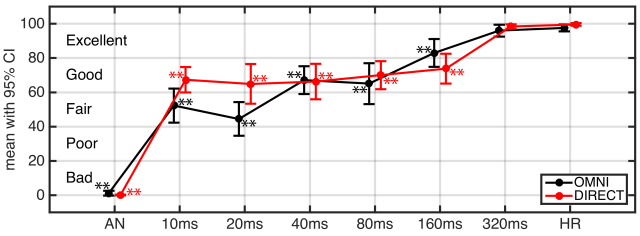

Abbildung 3: Bewertete Ähnlichkeit LEA, KVB-Saal, omnidirektionale (schwarz) vs. gerichtete (rot) Quelle; signifikant unterschiedlich zur Referenz: ** ($\alpha = 0.01$)

Experiment 2 (LEB)

Für den kleinen Sendesaal zeigt sich bei Verwendung der omnidirektionalen Schallquelle ein ähnlicher Verlauf der Bewertungen (siehe Abbildung 4) wie im Experiment 1 (LEA). So steigen die Bewertungen tendenziell mit der Überblendzeit in den statischen Nachhall an und liegen auch hier bereits ab einem Zeitpunkt von 5 ms im Bereich des Attributs *good*. Die Stimuli wurden ebenfalls bis zu dem Zeitpunkt von 160 ms signifikant schlechter im Vergleich zur Referenz bewertet. Im Gegensatz zum Versuch LEA wurden alle Stimuli sogar bei Verwendung eines Signifikanzniveaus von 1 % signifikant schlechter als die Referenz bewertet (LEA nur 5 ms, 10 ms und 20 ms signifikant bei $\alpha = 0.01$). Liegen die Mittelwerte der Bewertungen der Versuche LEA und LEB für die Zeitpunkte 5 ms, 10 ms und 20 ms noch recht nah beieinander, so zeigen sich für die Zeitpunkte 40 ms, 80 ms und 160 ms größere Unterschiede zwischen den Versuchen. Dabei fällt die Bewertung der Stimuli ohne Direktschall (LEB) signifikant schlechter aus im Vergleich zur Auralisation mit Direktschall (LEA), siehe Tabelle 5.
Bei Verwendung der gerichteten Schallquelle im kleinen Sendesaal bestehen zwischen den beiden Experimenten größere Unterschiede. Alle untersuchten Überblendzeitpunkte in den statischen Nachhall wurden ohne Direktschall selbst bei einem Signifikanzniveau von 1 % signifikant schlechter bewertet im Vergleich zur Referenz. Es zeigt sich zwar für beide Versuche ein ähnlicher Verlauf, jedoch liegen alle Mittelwerte der Bewertungen des Versuchs LEB unterhalb der des Versuchs LEA. Stellt man wiederum diese Bewertungen bei identischer Übergangszeit direkt gegenüber, so wurden die Stimuli ohne Direktschall (LEB) bei allen Zeitpunkten bis auf die früheste Überblendzeit von 5 ms signifikant schlechter bewertet als die Stimuli mit Direktschall (LEA).

Quelle	1	2	3	4	5	6
omni.	0.9	0.37	0.38	0.008	0.001	0.002
gerichtet	0.1	0.004	0.04	0.02	< 0.001	0.002

Tabelle 5: Vergleich der Bewertungen LEA vs. LEB (kleiner Sendesaal); Signifikanz-Werte aus Zweistichproben-t-Tests je Übergangszeitpunkt 1-6

Fazit

In einer perzeptiven Untersuchung mittels dynamischer Binauralsynthese wurden verschiedene Überblendzeiten in einen statischen Nachhall auralisiert und deren

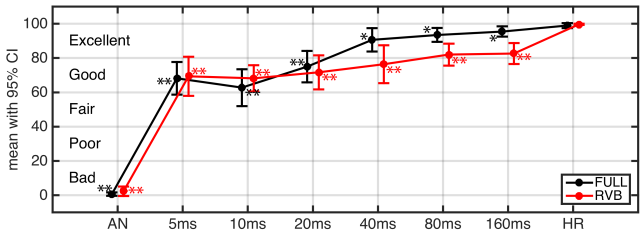

Abbildung 4: Bewertete Ähnlichkeit LEA (schwarz) vs. LEB (rot), kleiner Sendesaal, omnidirektionale Quelle; signifikant unterschiedlich zur Referenz: * ($\alpha = 0.05$) bzw. ** ($\alpha = 0.01$)

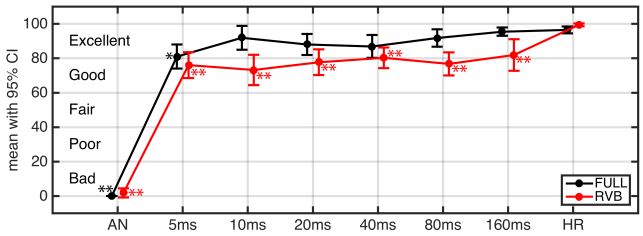

Abbildung 5: Bewertete Ähnlichkeit LEA (schwarz) vs. LEB (rot), kleiner Sendesaal, gerichtete Quelle; signifikant unterschiedlich zur Referenz: * ($\alpha = 0.05$) bzw. ** ($\alpha = 0.01$)

Unterschiede im Vergleich zur voll-dynamischen Repräsentation bewertet. Desweiteren wurden dabei unterschiedliche Richtcharakteristika (omnidirektional und gerichtet) als auch Auralisierungen mit und ohne Direktschall miteinander verglichen.

Tendenziell zeigt sich eine mit der Übergangszeit ansteigende Bewertung, die Stimuli wurden zudem alle mit dem Attribut *fair* oder besser bewertet (CQS > 40). Dies ist insbesondere bei frühen Übergangszeiten überraschend, da hier mitunter durch die Modifizierung der BRIR-Datensätze direkt in erste Reflexionen eingegriffen wurde. Desweiteren zeigt sich, dass teilweise auch hohe Zeitpunkte (z.B. kleiner Sendesaal bei 160 ms) signifikant unterschiedlich zur Referenz bewertet wurden, d.h. den Probanden war eine Unterscheidung der Stimuli weiterhin möglich. Dies spricht für ein noch nicht diffuses Schallfeld und deckt sich mit vorhergegangen Versuchen, die die vorliegende Untersuchung motivierten. Hier finden sich klare Differenzen zu Untersuchungen aus der Literatur, die in der Regel deutlich frühere perzeptive Mixing Times zeigen.

Eine Maskierungswirkung des Direktschalls konnte gezeigt werden: Zum einen wurden bei der Auralisierung ohne Direktschall in allen Konstellationen signifikante Unterschiede zur Referenz festgestellt, während mit Direktschall mitunter keine Unterscheidung mehr möglich war. Zum anderen wurden die Stimuli ohne Direktschall in der Regel signifikant unterschiedlicher bewertet als die entsprechenden Stimuli mit Direktschall. Dies spricht für eine Maskierung durch den Direktschall. Aufgrund der Maskierung werden Veränderungen im diffusen Nachhall teilweise unhörbar. Somit könnte das Verhältnis zwischen Direkt- und Diffusschall ein durchaus relevanter Faktor für die Bestimmung der perzeptiven Mixing Time sein.

Insbesondere bei frühen Übergangszeiten in einen statischen Nachhall konnten zudem signifikante Unterschiede zwischen den Richtcharakteristika nachgewiesen werden. Dabei wurde im Vergleich zur Referenz die gerichtete Schallquelle tendenziell ähnlicher als die omnidirektionale Schallquelle bewertet. Auch dies spricht für eine hohe Relevanz des Verhältnisses von Direkt- zu Diffusanteil. Bei der gerichteten Quelle dominiert der Direktschall, was möglicherweise zu einer stärkeren Maskierungswirkung führt.

Förderung

Diese Forschungsaktivitäten werden vom Bundesministerium für Bildung und Forschung (BMBF) im Programm *Forschung an Fachhochschulen* finanziert, Kennzeichen 03FH005I3 - MoNRa. Wir bedanken uns für die Unterstützung.

Literatur

[1] A. Lindau, L. Kosanke, and S. Weinzierl, “Perceptual evaluation of model-and signal-based predictors of the mixing time in binaural room impulse responses,“ in *Journal of the Audio Engineering Society*, vol. 60, no. 11, pp. 887-898, 2012.

[2] C. Pörschmann, A. Zebisch, “Psychoakustische Untersuchungen zu synthetischem diffusen Nachhall,“ in *Proc. of the VDT International Convention*, Cologne, 2012.

[3] L. Cremer, and H. A. Müller, “Die wissenschaftlichen Grundlagen der Raumakustik Bd. 1: Geometrische Raumakustik, Statistische Raumakustik, Psychologische Raumakustik,“ 2nd ed., Stuttgart, 1978.

[4] P. Rubak, and L. G. Johansen,“Artificial Reverberation Based on a Pseudo-Random Impulse Response II,“ presented at *106th Convention of the Audio Engineering Society*, preprint 4900, 1999.

[5] T. Hidaka, Y. Yamada, and T. Nakagawa,“A New Definition of Boundary Point between Early Reflections and late Reverberation in Room Impulse Responses,“ in *J. Acoust. S. Am.*, vol. 122, No. 1, pp. 326-332, 2007.

[6] P. Stade, B. Bernschütz, and M. Rühl,“A Spatial Audio Impulse Response Compilation Captured at the WDR Broadcast Studios,“ in *Proc. of the VDT International Convention*, Cologne, 2012.

[7] J. Ahrens, M. Geier, and S. Spors, “The soundscape renderer: A unified spatial audio reproduction framework for arbitrary rendering methods,“ in *Proc. of the Audio Engineering Society Convention 124*, 2008.

[8] B. Bernschütz, A. Vazquez, C. Pörschmann, and J. Arend, “Binaural Reproduction of Plane Waves With Reduced Modal Order, “ in *Acta Acustica united with Acustica*, vol. 100, no. 5, pp. 972-983, Sep. 2014.

[9] ITU, “Method for the subjective assessment of intermediate quality level of coding systems,“ in *Tech. Rept. Rec. ITU-R BS.1534-2*, 2014.

[10] A. Vazquez, “Scale - Conduction Psychacoustic Experiments with Dynamic Binaural Synthesis ,“ in *Proc. of the DAGA 2015*, Nuremberg, 2015.

[11] Y. Hochberg, “A sharper Bonferroni procedure for multiple tests of significance, “ in *Biometrika 75*, 800-2, 1988.