Kann man die Form eines Konzertsaales hören? Ein audiovisueller Test in simulierten 3D Umgebungen

Jakob Greif, David Ackermann, Omid Kokabi, Stefan Weinzierl
TU Berlin, Fachgebiet Audiokommunikation, Deutschland, Email: jakobgreif91@gmail.com

Einleitung

Den klassischen architektonischen Typen musikalischer Aufführungsräume wird traditionell eine eigene akustische Signatur zugeschrieben, die u.a. auf ein charakteristisches Muster früher Reflexionen zurückgeführt wird. Dabei ist bekannt, dass frühe Schallreflexionen aus dem Raum das akustische Perzept auf vielfältige Weise beeinflussen können. Sie können etwa die wahrgenommene Ausdehnung von Schallquellen erhöhen [1], und sie können ausgewertet werden, um die Richtung und den Abstand zu Wandbegrenzungsflächen zu schätzen [2]. Inwieweit es allerdings ein komplexes und im Einzelfall auch individuell unterschiedliches Muster früher Reflexionen ermöglicht, die architektonische Form von Räumen zu bestimmen, ist bisher eher anekdotisch beschrieben als empirisch belegt, zumal die Fähigkeiten zur kognitiven Analyse von Reflexionsmustern bei verschiedenen Personen unterschiedlich gut ausgeprägt zu sein scheint [3]. Wie einfach sich also ein „Weinberg“, ein „Fächer“, ein „Hufeisen“ oder eine „Schuhschachtel“ akustisch tatsächlich identifizieren lässt, wenn andere Parameter wie das Volumen, die Nachhallzeit oder der Streugrad der Wände identisch sind, wurde durch ein Experiment untersucht. Hierbei wurde ein stereoskopisches Display benutzt, um verschiedene Konzertsäle zu visualisieren, während die Akustik der Säle durch dynamische Binauralsynthese auralisiert wurde.

Methode

Für den Hörversuch wurden vier Konzertsäle vom Typus „Weinberg“, „Fächer“, „Hufeisen“ und „Schuhschachtel“ digital modelliert (siehe Abb. 1). Bei der Ausgestaltung der Säle wurden verschiedene reale Räume als Vorlage genommen. Einerseits galt es hierbei, die für den jeweiligen Typus konstitutiven architektonischen Elemente im Modell abzubilden. Andererseits sollten keine Elemente modelliert werden, die nur für einen bestimmten Saal charakteristisch sind. So wurden etwa die in den bekannten Schuhschachtel-Sälen fast immer vorhandenen Elemente wie überhängende Balkone oder die Orgel hinter der Bühne modelliert, während Säulen und die Feinstruktur der Seitenwänden nicht spezifiziert wurden (vgl. Abb. 1). Die Raumakustik der Säle wurde für diese Modelle mit dem hybriden Ray Tracing Tool RAVEN [4] simuliert. Dabei wurden frühe Reflexionen bis zur 3. Ordnung mit einem Spiegelschallquellen-Modell berechnet. Der spätere Nachhall wurde durch stochastisches Ray Tracing mit je 400k Partikeln für 31 Terzbänder berechnet. Eine Schallquelle mit der Richtcharakteristik eines Sprechers wurde auf der Bühne platziert, da männliche Sprache als Quellsignal ausgewählt wurde. In den Sälen wurde eine Publikumsfläche mit den akustischen Eigenschaften von zur Hälfte besetzten, gepolsterten Stuhlreihen definiert, so dass die Akustik einem Mittelwert zwischen besetztem und unbesetztem Zustand entsprach. Abgesehen von der Publikumsfläche und der Fläche der Orgel wurden alle anderen Oberflächen mit einem homogenen Restschall-Absorptionsgrad belegt, der für jeden Raumtyp und jede Kombination der restlichen akustischen Bedingungen individuell berechnet wurde, um die gewünschte Nachhallzeit zu erreichen.

Die Hörpositionen befanden sich in 10 m Abstand zur Schallquelle in den kleinen und 15 m in den großen Sälen. Dadurch wurde sichergestellt, dass sich alle Positionen außerhalb des Hallradius befinden. Mit den Außenohrübertragungsfunktionen (HRTFs) des FABIAN Kunstkopfes [5] wurden Datensätze von binauralen Raumimpulsantworten (BRIRs) für verschiedene Kopforientierungen erzeugt, in denen die Eigenschaften der akustischen Szene kodiert waren.

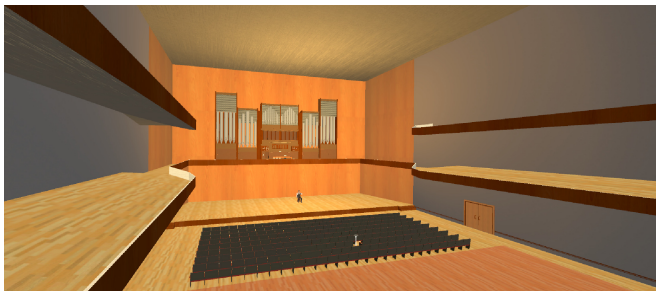
Um den Einfluss verschiedener akustischer Bedingungen zu untersuchen, wurde das Volumen, die Nachhallzeit und der Streugrad der Oberflächen in jeweils zwei Stufen variiert (siehe Tab. 1). Der Streugrad wurde theoretisch nach [6] für zufällig strukturierte Flächen mit Strukturtiefen von 1 cm und 5 cm berechnet, woraus sich jeweils eine sigmoide Funktion $\gamma(f)$ für den Streugrad ergibt, die bei Übergangsfrequenzen von 650 Hz bzw. 3.5 kHz von $\gamma(f) = 0,1$ auf $\gamma(f) = 0,9$ ansteigt. Für jeden Raumtypus wurden somit BRIR-Datensätze für alle acht Kombinationen der drei akustischen Bedingungen berechnet.

Tabelle 1: Unabhängige Variablen, deren Einfluss auf die Erkennung der vier Raumtypen untersucht wurde.

Training	Ja	Nein
Nachhallzeit [s]	1,5	2,5
Volumen [m^3]	5.000	20.000
Streugrad	klein	groß

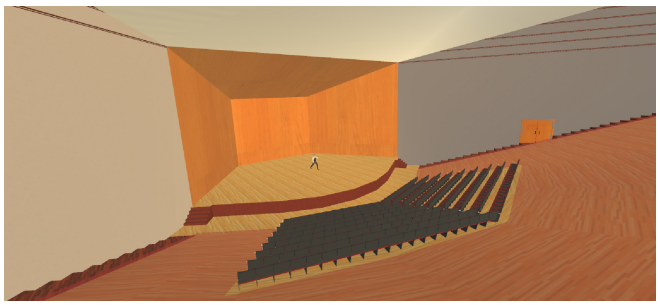
Die Räume wurden mithilfe des SoundScape Renderers (SSR, [7]) über Kopfhörer (Beyerdynamic DT770 Pro) dynamisch binaural auralisiert, d.h. der Kopforientierung in einem Bereich von ±180° azimutal und [−30°..., 60°] vertikal nachgeführt. Die Übertragungsfunktion des Kopfhörers wurde durch regularisierte Inversion digital entzerrt. Die Kopforientierung wurde über einen Head-Tracker (Polhemus Patriot FCC Class B) gemessen. Die visuelle Testumgebung wurde über ein stereoskopisches Display (Oculus Rift) wiedergegeben. Um einen ausreichend glaubwürdigen Eindruck eines realen Konzertsaales zu vermitteln wurden die Modelle mit Texturen,

(a) Weinberg



(b) Schuhschachtel

(c) Hufeisen



(d) Fächer

Abbildung 1: Darstellung der Konzertsäle mit Texturierung und graphischen Details, wie sie in der virtuellen Testumgebung dargestellt wurden.

Türen, wichtigen Objekten wie der Orgel und Stuhlreihen grafisch aufgearbeitet (s. Abb. 1).

Für den Hörtest wurde ein 4-AFC-Verfahren gewählt, bei dem Probanden je vier Auralisationen der verschiedenen Raumtypen mit dem visuellen Eindruck eines Saales vergleichen und die korrekte Auralisation auswählen sollten. Das Auswahlmenü aus der Perspektive der Teilnehmenden mit einer Darstellung des virtuellen „Weinbergs“ ist in Abbildung 2 dargestellt. In 32 randomisierten Wiederholungen wurde jeder der vier Konzertsäle mit allen acht Kombinationen der akustischen Bedingungen jeweils einmal als Zielreiz präsentiert. Als Quellsignal wurde die nachhallfreie Aufnahme eines professionellen Schauspielers verwendet [8], die geeignet war, auch in größeren Sälen und größeren Entfernungen von der Quelle eine gute Sprachverständlichkeit und -intensität zu gewährleisten. In einem informellen Vorversuch hatte sich das Sprachsignal, im Vergleich zu verschiedenen Musiksignalen, aufgrund seiner Mischung von transienten Signalanteilen und Pausen im Signal, als besonders geeignet für die Erkennung der akustischen Signatur des Raums erwiesen. Um den Einfluss eines vorangehenden Trainings auf die Erkennungsleistung der Hörer*innen zu untersuchen, erhielt die Hälfte der Teilnehmenden vor dem Test ein freies Training, bei dem alle Säle gleichzeitig visuell und akustisch verglichen werden konnten. Die Teilnehmenden hatten 15 min Zeit, um Cues für die raumakustische Charakteristik der Säle zu identifizieren und diese unter den verschiedenen, akustischen Testbedingungen zu vergleichen. Um Lernermüdung zu vermeiden konnten die Räume im Training nur mit geringem Streugrad angehört werden. Die andere Hälfte der Teilnehmenden erhielten vor dem Test lediglich visuelle Ansichten der Konzertsäle.

Die Ergebnisse des Hörversuchs wurden mithilfe eines generalisierten linearen gemischten Modells (GLMM) analysiert, um den Einfluss der verschachtelten Testbedingungen auf die binäre Variable „korrekte Auswahl“ abzubilden. Dabei werden sowohl konstante Haupteffekte (fixed effects) als auch Unterschiede zwischen den Versuchspersonen (Zufallseffekte, random effects) modelliert. Die Modellgüte wurde über das Bestimmtheitsmaß (pseudo$-R^2$) gemessen [10]. Dabei beschreibt R^2_{marginal} den Anteil der von den konstanten Effekten erklärten Varianz, $R^2_{\text{conditional}}$ erfasst zusätzlich die Zufallseffekte und damit die Güte des Gesamtmodells.

Abbildung 2: Perspektive der Versuchsteilnehmer im virtuellen Weinberg-Saal. Das Auswahlmenü befand sich im Sichtfeld und konnte über einen virtuellen Pointer bedient werden. Die Hörposition befindet sich im vorderen Parkett, 1 m neben der Mittelachse des Konzertsaales.

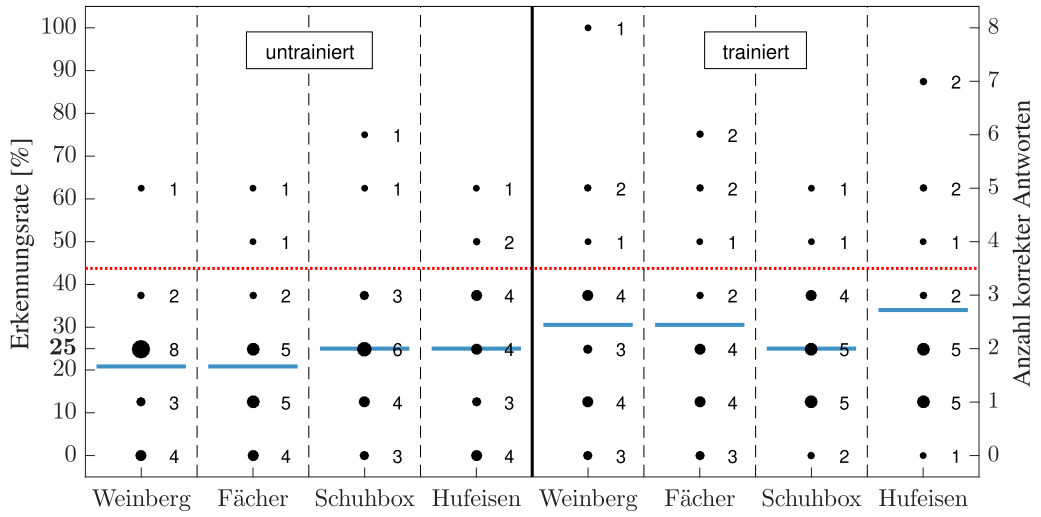


Abbildung 3: Erkennungsrate und absolute Anzahl korrekter Antworten aller untrainierten (links) und trainierten Probanden (rechts) für die verschiedenen Raumtypen. Mittelwerte sind als blaue Striche markiert. Die rote Linie markiert die Grenze, ab der die Erkennungsrate einer VP signifikant oberhalb der Ratewahrscheinlichkeit (25 %) liegt.

An der Untersuchung nahmen 9 weibliche und 27 männliche Testpersonen im Alter von 22–53 Jahren teil. Die meisten Teilnehmer*innen gaben in einer Selbstauskunft an, fachliches Wissen über Raumakustik zu besitzen; 6 Personen gaben an, regelmäßig klassische Konzerte zu besuchen.

Ergebnisse

Abbildung 3 zeigt die individuellen Trefferquoten aller Probanden sowie die Mittelwerte für die einzelnen Raumtypen. Bei Betrachtung der Trefferquoten der Probandengruppe als Ganzes konnte keine signifikante Abweichung von der Ratewahrscheinlichkeit (25 %) festgestellt werden, da berechnete Konfidenzintervalle (nicht angezeigt) stets auch den Bereich der Ratewahrscheinlichkeit einschließen. Bei individueller Betrachtung zeigt sich jedoch, dass ein Sechstel der Probanden die Raumtypen signifikant oberhalb der Ratewahrscheinlichkeit identifizieren konnte. Drei trainierte Probanden konnten, gemittelt über alle Raumtypen, eine Erkennungsrate von 60 % erreichen.

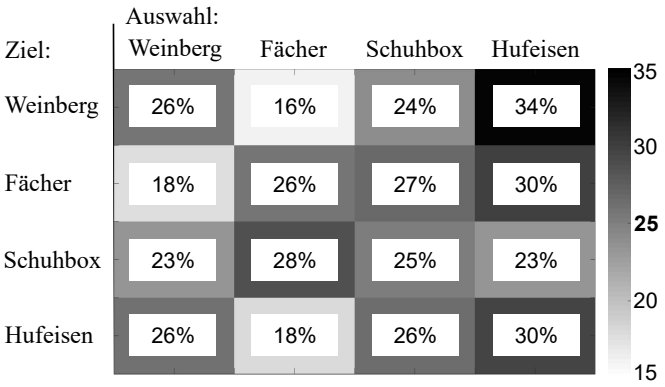


Abbildung 4: Konfusionsmatrix für die Antworthäufigkeiten bei der Identifikation der vier Raumtypen.

In der Konfusionsmatrix der Antworten (Abb. 4) zeigt sich kein klares und konsistentes Muster, welche Raumtypen mit welchen anderen Raumtypen bevorzugt verwechselt wurden, zumal alle Antworten nahe an der Ratewahrscheinlichkeit liegen.

Den Einfluss der verschiedenen Testbedingungen auf die Erkennung der Raumtypen zeigt das GLMM-Modell, wobei in Tabelle 2 sowohl die Haupteffekte als auch die signifikanten Interaktionseffekte ($p < 0,05$) dargestellt sind.

Tabelle 2: Ergebnisse des gemischten linearen Modells (GLMM) mit den akustischen Faktoren Volumen, Nachhallzeit und Streugrad, dem Training und den Raumtypen (Dummykodierung) als unabhängige Variablen. Der Koeffizient beschreibt Größe und Richtung des Einflusses der Variablen.

Haupteffekte	Koeffizient	p-Wert
Nachhallzeit	-0,51	**0,04**
Streugrad	0,1	0,7
Volumen	0,28	0,32
Training	0,53	**< 0,01**
Fächer	-0,21	0,47
Schuhschachtel	-0,05	0,86
Hufeisen	-0,14	0,62
Interaktionseffekte		
Streugrad×Training	-0,51	**< 0,01**
Nachhallzeit×Schuhschachtel	0,71	**0,041**
Nachhallzeit×Hufeisen	0,92	**< 0,01**

Die Modellgüte beläuft sich auf $R^2_{\text{marginal}} = 0,08$ und $R^2_{\text{conditional}} = 0,15$. Die vier unabhängigen Variablen (Tab. 1) erklären also nur 8 % der Varianz in der Erkennungsleistung, zusammen mit der individuellen Leistungsfähigkeit der Probanden werden 15 % der Varianz erklärt. Die geringen Werte sind nicht überraschend, da nur wenige Teilnehmer den Test besser als mit Ratewahrscheinlichkeit absolvieren konnten.

Signifikante Haupteffekte liegen für die Testbedingungen *Nachhallzeit* und den Faktor *Training* vor. Das Vorzeichen des Koeffizienten deutet darauf hin, dass die Teilnehmer die Raumformen besser identifizieren konnten, wenn eine geringere Nachhallzeit vorlag und wenn ein vorangehendes Training absolviert wurde. In ihrem Einfluss auf das Ergebnis liegen beide Effekte in der gleichen Größenordnung. Keinen Einfluss hatten das Volumen und der Streugrad der Räume.

Ein hoch signifikanter Interaktionseffekt wurde zwischen den Faktoren *Training* und *Streugrad* gefunden. Teilnehmer konnten die Konzertsäle mit geringer Wandstreuung dann besser identifizieren, wenn sie zuvor ein Training absolviert haben. Dieser Effekt muss allerdings als Artefakt durch das Training gedeutet werden, denn dort wurden ausschließlich die Säle mit geringer Streuung verwendet. Zum einen hatten die Probanden diese Stimuli nachfolgend wohl noch besser in Erinnerung, zum anderen konnten sie den dadurch erzielten Trainingseffekt offensichtlich nicht auf die gleichen Konzertsäle mit höherer Streuung der Wände übertragen. Signifikante Interaktionseffekte konnten auch zwischen dem Faktor *Nachhallzeit* und den Raumtypen *Schuhschachtel* und *Hufeisen* gefunden werden. Sie weisen darauf hin, dass der Verlust an Erkennbarkeit mit steigendem Nachhall (s. Haupteffekte) bei Hufeisen- und Schuhschachtel-Räumen nicht so groß war wie bei den (für die Dummykodierung als Referenz verwendeten) Weinberg-Sälen.

Diskussion

In dieser Arbeit wurde durch einen Hörversuch in audiovisuellen virtuellen Umgebungen untersucht, inwieweit die raumakustische Signatur von vier klassischen Raumtypen für Konzertsäle auditiv identifiziert werden kann, und in welchem Umfang diese Erkennung vom Volumen, von der Nachhallzeit der Räume und vom Streugrad der Wände abhängt. Um es zu ermöglichen, diese Parameter unabhängig von der Geometrie des Raums zu verändern, wurde nicht mit gebauten Räumen, sondern mit raumakustischen Simulationen gearbeitet. Da es im Versuch nicht darauf ankam, spezifische Räume klanglich nachzubilden, sondern durch Raumsimulation und dynamische Binauralsynthese ein akustisch plausibles räumliches Gesamtbild zu schaffen [9], und da die akustische Signatur der verschiedenen architektonischen Typen mutmaßlich durch starke, frühe Reflexionen an großen Oberflächen geschaffen wird, welche durch eine geometrische Simulation gut simuliert werden können, gehen wir davon aus, dass die Validität der Ergebnisse dadurch nicht eingeschränkt ist.

Im Ergebnis waren die Probanden nur in einem Sechstel der Testdurchläufe in der Lage, den Klang der Räume dem optisch präsentierten Raumtypus mit einer Erkennungsrate, die signifikant oberhalb der Ratewahrscheinlichkeit lag, korrekt zuzuordnen. Einen signifikanten Einfluss auf die Erkennungsrate hatte einerseits die Nachhallzeit der Räume und andererseits das vor dem Versuch absolvierte Training der Teilnehmer.

Bei kürzerer Nachhallzeit konnten die Raumtypen besser erkannt werden konnten, vermutlich weil die durch frühe Reflexionen definierte Signatur der Raumtypen stärker hervortritt, wenn der späte Nachhall schneller an Energie verliert. Dieser Effekt war bei Weinberg-Sälen stärker ausgeprägt als bei Schuhschachtel- und Hufeisen-Sälen. Nach informellen Rückmeldungen der Probanden war gerade bei längerem Nachhall bei Schuhschachtel- und Hufeisen-Sälen die – gegenüber den Weinbergsälen – größere Umhüllung durch rückwärtig einfallende Schallanteile spürbar.

Einen signifikanten Einfluss auf die Erkennungsrate hatte auch das unmittelbar vor dem Versuch absolvierte Training. Keinen Einfluss im Modell hatte allerdings die von den Teilnehmenden selbst angegebene raumakustische Expertise, die bei der Analyse probeweise als Prädiktor in das lineare Modell eingefügt wurde. Somit scheint der beobachtete Effekt eher das Resultat eines spezifischen und zeitnah zum Versuch absolvierten Trainings zu sein als das Ergebnis von allgemeiner raumakustischer Erfahrung.

Auch unter günstigen Bedingungen, d.h. mit einer für Konzertsäle kurzen Nachhallzeit von 1,5 s und mit einem spezifischen vorangehenden Training, erzielten die Teilnehmer im Mittel gegenüber einer Ratewahrscheinlichkeit von 25% jedoch nur eine Erkennungsrate von 28.8 %, und einzelne trainierte Probanden erzielten eine Erkennungsrate von 60 %. Insofern muss man festhalten, dass der „Charakter“ von architektonischen Raumtypen offensichtlich akustisch insgesamt nicht so deutlich ausgeprägt und so klar identifizierbar ist wie es häufig angenommen wird.

Danksagung

Für die Beratung bei der Erstellung der Raummodelle bedanken wir uns bei Eckhard Kahle (Kahle Acoustics), für statistische Beratung bei Steffen Lepa.

Literatur

[1] Blau, M.: Correlation of apparent source width with objective measures in synthetic sound fields, Acta Acustica united with Acustica 90 (2004), 720–730.

[2] Thaler, L.: Echolocation in humans: an overview. Cognitive Science 7.6 (2016), 382–393.

[3] Schenkman, Bo N. und Nilsson, M. E: Human echolocation: Blind and sighted persons' ability to detect sounds recorded in the presence of a reflecting object, Perception 39 (2010), 483–501.

[4] Schröder, D.: Physically based real-time auralization of interactive virtual environments. Dissertation RWTH Aachen, Berlin 2011.

[5] Brinkmann, F., Lindau, A., Weinzierl, S., van de Par, S., Müller-Trapet, M., Opdam, R. und Vorländer, M.: A High Resolution and Full-Spherical Head-Related Transfer Function Database for Different Head-Above-Torso Orientations. J. Audio Eng. Soc. 65 (2017), 841–848.

[6] Embrechts, J.-J., Archambeau, D. und Stan, G.-B.: Determination of the scattering coefficient of random rough diffusing surfaces for room acoustics applications. Acta Acustica united with Acustica 87 (2001), 482–494.

[7] Ahrens, J.: The soundscape renderer: A unified spatial audio reproduction framework for arbitrary rendering methods. In 124th AES Convention (2008), 2203–2503.

[8] Böhm, C., Fiedler, F. und Weinzierl, S.: An Anechoic Recording of Cicero's 3rd Cataline Oration: Italian, Latin and German, DOI: 10.14279/depositonce-8536 (2019).

[9] Lindau, A. und Weinzierl, S.: Assessing the plausibility of virtual acoustic environments. Acta Acustica united with Acustica 98.5 (2012), 804–810.

[10] Nakagawa, S. und Schielzeth, H.: A general and simple method for obtaining R2 from generalized linear mixed-effects models. Methods in ecology and evolution 4 (2013), 133–142.