Abschlussbericht

zum DFG-Forschungsvorhaben

Immersive binaurale Kommunikation | Modelle und Algorithmen

Autoren: Sebastian Nagel, Tobias Kabzinski, Erik Fleischhauer, Christiane Antweiler, Peter Jax

Institut für Kommunikationssysteme (IKS)
RWTH Aachen University
Muffeter Weg 3a
52074 Aachen

Erscheinungsort: Aachen
Erscheinungsjahr: 2026

Verfügbar über das institutionelle Repositorium der RWTH Aachen University
DOI: 10.18154/RWTH-2026-00152

Inhalt

ABSCHLUSSBERICHT

1 Allgemeine Angaben

DFG-Geschäftszeichen: JA 2746/3-1
Projektnummer: 509806277
Titel des Projekts: Immersive binaurale Kommunikation | Modelle und Algorithmen
Name des Antragstellers: Prof. Dr.-Ing. Peter Jax
Dienstanschrift: Institut für Kommunikationssysteme, 52056 Aachen
Name der Mitverantwortlichen: Dr.-Ing. Christiane Antweiler
Berichtszeitraum (gesamte Förderdauer): 04/2023–03/2025

2 Zusammenfassung/Summary

Zusammenfassung

Das Projekt behandelt die Telekommunikation mit räumlichen Audiosignalen. Als Anwendungsszenario wird eine „hybride“ Telekonferenz betrachtet, bei der sich eine Gruppe von Personen in einem realen Raum trifft und eine weitere Person aus der Ferne teilnimmt. Mit den derzeit üblichen digitalen Sprachkommunikationssystemen kann ein solches Szenario für die ferne Person zu einer ermüdenden und unnatürlichen Erfahrung führen, da die fehlende räumliche Akustik und die Notwendigkeit, visuelle und auditive Informationen gleichzeitig zu verarbeiten, zur sogenannten „Zoom-Fatigue“ beitragen. Dem wirkt das Projekt entgegen: Durch Aufzeichnung und Übertragung räumlicher 3D-Audiosignale aus dem Besprechungsraum kann die ferne Person ihre zum natürlichen Sprachverstehen erlernten kognitiven Prozess einsetzen, um intuitiv die Sprecher im Raum zu unterscheiden, sich gezielt auf einzelne Sprecher zu konzentrieren und mögliche Nebengeräusche auszublenden.

Mit binauraler Aufnahmetechnik werden zur Aufzeichnung der 3D-Audiosignale lediglich zwei Mikrofone auf den Seiten einer Kugel oder an den Ohren eines nachgebildeten Kopfes benötigt. Der technische Aufwand für Aufzeichnung und Übertragung ist so im Vergleich zu alternativen Techniken viel geringer. Die im Projekt entwickelten Algorithmen und Modelle überwinden bisher bestehende grundlegende Einschränkungen bei der Wiedergabe derartig aufgezeichneter Signale. Sie ermöglichen die räumliche Wiedergabe über Lautsprecher, sodass der Zuhörer für die räumliche Wahrnehmung keine Kopfhörer tragen muss. Dabei ermöglichen sie dem Zuhörer, den Kopf in Richtung einzelner Gesprächspartner zu drehen und die einzelnen Sprecher an feststehenden Positionen im Raum wahrzunehmen.

Die Ergebnisse des Projekts führen dazu, dass die aus der Ferne zugeschaltete Person besser akustisch in den Besprechungsraum hineinversetzt wird (Immersion), während gleichzeitig der technische Aufwand für Aufzeichnung und Übertragung niedrig gehalten wird. Auf diese Weise trägt das Projekt grundlegend dazu bei, komfortablere Telekonferenzen zu ermöglichen, die eine realistische Raumakustik wiedergeben. Neben der Schonung von zeitlichen und finanziellen Ressourcen könnten sich dadurch auch positive Effekte für Klima und Umwelt ergeben. Die Ergebnisse setzen außerdem grundlegende Impulse für die Wissenschaft, die sich auch auf andere Forschungsgebiete übertragen lassen, z.B. auf die mehrkanalige Echokompensation oder Hörgeräte.

Summary

The project deals with spatial audio telecommunication. It considers the scenario of a „hybrid“ teleconference, where a group of people meets in a real room and another person participates remotely. With current communication systems, this experience is tiring and unnatural for the remote person. The missing spatial acoustics and the need to process visual and auditive information simultaneously lead to so-called „Zoom fatigue“. The project addresses these issues: By recording and transmitting spatial audio signals from the meeting room, the remote person can use cognitive processes to intuitively differentiate speakers, focus on individual speakers and suppress background noise.

Binaural recording technology is an attractive option in this scenario. It requires only two microphones on the sides of a sphere or on the ears of a replica head for 3D audio recording. The technical effort for recording and transmission is therefore much lower compared to alternative techniques. The algorithms and models developed in the project overcome previously existing fundamental limitations in the reproduction of such recorded signals. They enable the spatial reproduction of binaurally recorded signals via loudspeakers, so that the listener does not have to wear headphones for a spatial perception. They also enable the listener to turn their head in the direction of individual conversation partners and to perceive the individual speakers at fixed positions in the room.

As a result of the project, the remote person can be acoustically immersed into the meeting room, while at the same time keeping the technical effort for recording and transmission low. In this way, the project contributes fundamentally to enabling more comfortable teleconferences, which reproduce realistic room acoustics. In addition to saving time and financial resources, this could also have positive effects on the climate and environment. The results also provide fundamental impulses for science that can be transferred to other research areas, for example multi-channel acoustic echo cancellation or hearing aids.

3 Wissenschaftlicher Arbeits- und Ergebnisbericht

3.1 Ausgangsfragen und Zielsetzung des Projekts

Das betrachtete Anwendungsszenario im Projekt ist die räumliche Audio-Kommunikation in einer „hybriden“ Telekonferenz, wie in Abbildung 1 dargestellt. Die Nutzung binauraler Aufnahmetechnik bietet eine gute räumliche Wiedergabe bei geringem technischen Aufwand für Aufzeichnung und Übertragung. Bisher sind binaural aufgezeichnete Signale jedoch in ihren Anwendungsmöglichkeiten eingeschränkt. Zum einen leidet die räumliche Illusion, wenn der Zuhörer seinen Kopf bewegt. Da der Kunstkopf auf der Aufnahmeseite fest steht, hat der Zuhörer dann den Eindruck, dass sich die fernen Sprecher mit seinem Kopf mitbewegen, statt ortsstabil an festen Positionen im Raum zu stehen, wie in einer natürlichen Situation. Zum anderen wird die räumliche Wahrnehmung nur bei Kopfhörerwiedergabe hervorgerufen. Dies schließt das Ohr jedoch zu einem gewissen Grad ab, was insbesondere beim Sprechen häufig als unangenehm empfunden wird. Das Projekt zielt darauf, beide Einschränkungen zu überwinden, um die zugeschaltete Person akustisch in den Besprechungsraum hineinzuversetzen (Immersion).

Projektziele

In der natürlichen Sprachkommunikation dreht ein Zuhörer seinen Kopf in Richtung einzelner Gesprächspartner, deren Stimme dann von vorne wahrgenommen wird. An den Ohren des fernen Zu-

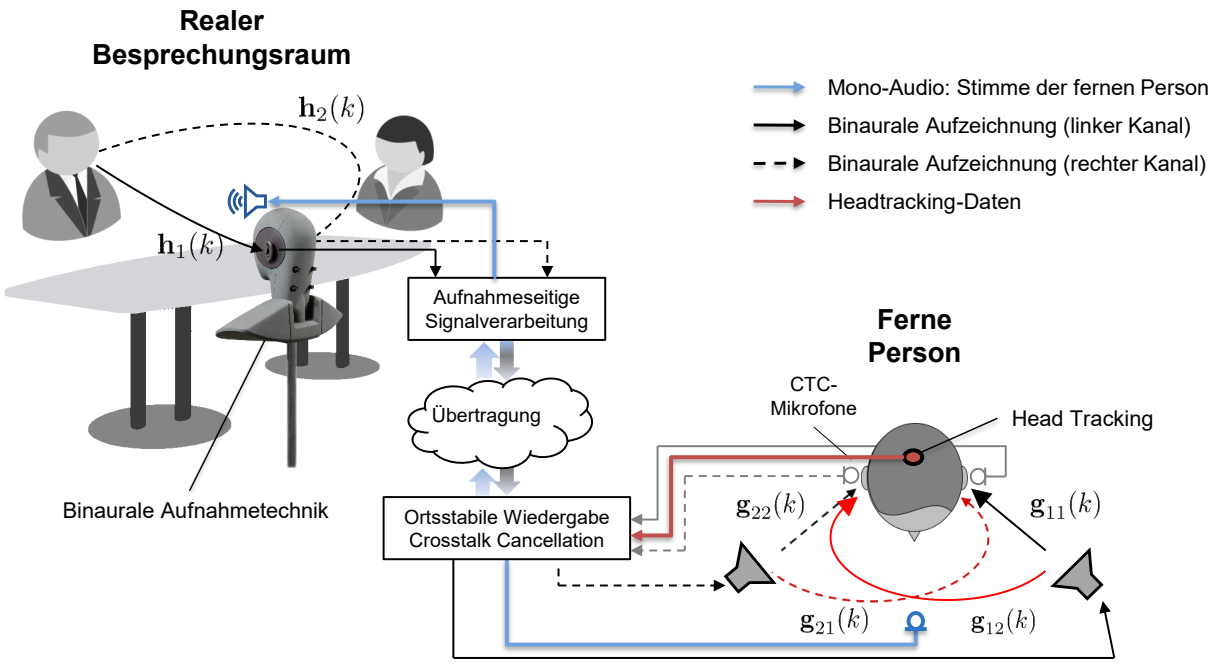

Abbildung 1: Binaurale Audio-Kommunikation. Links: Binaurale Aufnahme mittels Kunstkopf. Rechts: Wiedergabe der binaural aufgezeichneten Signale mittels Lautsprecher.

hörers muss dazu ein binaurales Signal erzeugt werden, dessen räumliche Eigenschaften sich nach diesen Kopfbewegungen richten. Es liegt jedoch ein aufgezeichnetes Signal vor, dessen Eigenschaften sich aus der Ausrichtung des Aufzeichnungssystems ergeben und das im Allgemeinen nicht zu den Bewegungen des Zuhörers passt.

Gegenstand des Projekts waren daher Algorithmen, die die räumlichen Eigenschaften binaural aufgezeichneter Signale wiedergabeseitig an gemessene Kopfbewegungen eines Zuhörers anpassen, und so eine plausible ortsstabile Wiedergabe erzielen. Bisherige Algorithmen eignen sich nicht für binaural aufgezeichnete Signale, sondern erfordern aufnahmeseitig eine räumlich dichtere Abtastung des Schallfelds.

Für den gewünschten räumlichen Höreindruck müssen die binauralen Signale an den Ohren des Zuhörers reproduziert werden. Üblicherweise geschieht dies über Kopfhörer. Zur Wiedergabe über Lautsprecher ist ein sogenanntes *Crosstalk Cancellation (CTC)*-Filternetzwerk erforderlich. Dieses kompensiert die akustischen Übertragungsfunktionen von den Lautsprechern zu den Ohren des Zuhörers, damit dort das gewünschte Binauralsignal reproduziert wird. Da sich die akustischen Übertragungsfunktion bei Kopfbewegungen ändern, müssen die CTC-Filter ebenfalls an Kopfbewegungen des Zuhörers angepasst werden.

Im Rahmen des Projekts wurde die *adaptive* CTC weiterentwickelt, bei der die zu kompensierenden akustischen Übertragungsfunktionen mit den Signalen ohrnaher Mikrofone identifiziert werden. An diese mehrkanalige Systemidentifikation werden hohe Anforderungen gestellt, da herkömmliche Methoden bei Kopfbewegung nicht hinreichend schnell auf die Änderungen reagieren. Eine neuartige Idee war es, wiederkehrende Muster in den Übertragungsfunktionen auszunutzen, um bei Kopfbewegung die Systemidentifikation zu beschleunigen. Die Lautsprecherwiedergabe mit CTC soll idealerweise das gleiche räumliche Abbild vermitteln wie ein Referenzsystem mit Kopfhörern.

Erfolgskriterien

Die Erfolgskriterien des Projekts waren wie folgt festgelegt:

1. Manipulation eines binaural aufgezeichneten Signals abhängig von der Kopfbewegung eines Zuhörers für eine plausible ortsstabile Wiedergabe,
2. ausreichend schnelle und robuste Identifikation der Übertragungsfunktionen zwischen Lautsprechern und Mikrofonen, sodass
3. die binaurale Wiedergabe über Lautsprecher mittels adaptiver CTC den gleichen räumliche Eindruck vermittelt wie ein Referenzsystem mit Kopfhörer und Head Tracker, sowie
4. Zusammenführung beider Ansätze in ein gemeinsames System zur ortsstabilen, lautsprecherbasierten Wiedergabe binaural aufgezeichneter Signale.

3.2 Beschreibung der durchgeführten Arbeiten und Ergebnisse

Im Folgenden werden Arbeiten und Ergebnisse im Hinblick auf die Erfolgskriterien dargestellt.

3.2.1 Proof of Concept auf Basis bestehender Methoden

Zunächst wurde ein *Proof of Concept (PoC)*-Gesamtsystem aufgebaut wie in [1] näher beschrieben. Dabei stand die Echtzeitfähigkeit im Vordergrund, sodass auf bestehende Algorithmen mit geringer rechentechnischer Komplexität und dementsprechend eingeschränkter Leistungsfähigkeit zurückgegriffen werden musste. Die Ergebnisse der Untersuchung [1] zeigen dennoch, dass es grundsätzlich möglich ist, mit dem vorgesehenen System die wichtigen binauralen Cues *Interaural Level Difference (ILD)* und *Interaural Time Difference (ITD)* an den Ohren des Zuhörers zu reproduzieren. Die systematische Untersuchung von Wechselwirkungen der in den Teilprojekten zu entwickelnden Algorithmen (ortsstabile Wiedergabe und adaptive CTC) zeigt auf, welche Anforderungen bestehen. Ein nennenswertes Ergebnis ist, dass die *Interaural Coherence (IC)* nur eingeschränkt reproduziert werden konnte. Dies war insbesondere während der Zuhörerbewegung der Fall und führte in informellen Hörtests zu einer fehlenden räumlichen Lokalisationsschärfe. Die frühzeitige Anfertigung eines PoC-Systems ermöglichte die Priorisierung der weiteren Arbeitsschritte.

3.2.2 Adaptive Wiedergabe binauraler Signale mittels Lautsprechern

Die Theorie der lautsprecherbasierten Wiedergabe von Binauralsignalen ist hinreichend bekannt und die Anwendung grundsätzlich praktikabel. Unter dieser Annahme stand die Entwicklung von Algorithmen zur Systemidentikation von zeitvarianten Systemen im Vordergrund. Beim Aufbau des PoC-Systems in [1] wurde auch festgestellt, dass ein relativ langer Teil der *Binaural Room Impulse Response (BRIR)* identifiziert werden muss, um CTC-Filter entwerfen zu können, die hinreichend viel Entzerrung im Direktpfad und Unterdrückung des Übersprechens im Kreuzpfad erzielen können. Wie genau die psychoakustischen Anforderungen für eine hinreichend plausible Wiedergabe sind, ist derzeit noch unklar. Die Verwendung von adaptiven Filtern mit im Vergleich zur vollständigen Raumimpulsantwort kurzen geschätzten Impulsantworten, z.B. nur entsprechend dem Direktschall, geht mit einem erhöhten effektiven Fehlersignal einher, das wiederum die Schätzung der kurzen Impulsantwort negativ beeinflusst. Da für eine gut generalisierende Modellierung von vollständigen Raumimpulsantworten wegen der großen Filterlänge große Datenmengen nötig sind und mindestens eine

Teilschätzung der frühen Reflektionen und des Nachhalls erforderlich erschienen, wurde der Fokus der Systemidentifikationsmethoden zunächst auf Methoden ohne explizites Vorwissen über die zu schätzenden Übertragungsfunktionen gesetzt.

Schnelle und robuste Systemidentifikation zeitvariabler Systeme

Aufbauend auf zustandsraumbasierten Schätzmethoden für Impulsantworten, insbesondere Kalman-Filter-basierten Methoden, wurde im Rahmen des Projekts in [2] eine Methode entwickelt, die nicht auf ein offline trainiertes Modell aufbaut, sondern zur Laufzeit sowohl Schätzungen der Systemimpulsantworten als auch der Zustandsraummodellparameter liefert. Dazu wird der *Expectation Maximization (EM)*-Algorithmus eingesetzt, was eine algorithmische Verzögerung nach sich zieht. Die Schätzung der Zustandsraumparameter erlaubt eine besseres Tracking der zeitlichen Änderungen der Systemimpulsantworten. Der in [2] entwickelte Ansatz erlaubt eine Vielzahl von Annahmen über die lokalen dynamischen Abhängigkeiten der Zustände, sodass z.B. eine Kopplung der Impulsantworten zum rechten und linken Ohrmikrofon modelliert werden kann. Dies ist in der vorliegenden Anwendung angebracht, da die Mikrofone sich nur (durch den Kopf des Zuhörers gekoppelt) gemeinsam bewegen können. Der Ansatz in [2] geht jedoch weit über die im Projekt betrachtete Anwendung hinaus und ermöglicht die Identifikation von zeitvarianten gekoppelten MIMO-Systemen in vielfältigen Fragestellungen. Besonders hervorzuheben ist hier das Potential für die Identifikation von *Head-Related Transfer Functions (HRTFs)*, da die zusätzliche Verzögerung dabei keine Einschränkung darstellt. Schnelle Messfahren könnten für die Entwicklung von Personalisierungsmodellen für HRTFs in der Zukunft eine wichtige Rolle spielen. Untersuchungen zeigen, dass unterschiedliche Kopplungsmodelle mit verschiedenen Rechenkomplexitäten auch mit unterschiedlicher Leistungsfähigkeit bei der Systemidentifikation einhergehen.

Parallel zur Entwicklung von Methoden, die ohne Vorwissen auskommen, wurden auch auf *a priori*-Wissen basierende Methoden zur Systemidentifikation von BRIRs untersucht. Im Rahmen des Projekts wurde eine auf Vorwissen über die zu schätzenden Impulsantworten beruhende Variante des EM-Algorithmus theoretisch beleuchtet, die eine potentiell verbesserte Impulsantwortschätzung bietet, wenn Trainingsdaten über die zu schätzenden Impulsantworten vorliegen[1]. Dazu werden die zu schätzenden Impulsantworten in einem im Vergleich zur Impulsantwortlänge niedrig-dimensionalen affinen Unterraum geschätzt, der über eine *Principle Component Analysis (PCA)* von Trainingsmaterial aufgestellt wird. Im Gegensatz zur Annahme, dass die zu identifzierenden Impulsantworten in einem affinen Unterraum liegen, kann auch angenommen werden, dass die Impulsantworten auf oder nahe einer niedrigdimensionalen Mannigfaltigkeit (engl. manifold) liegen. Mittels eines generativen statistischen Modells in Form des sog. *Variational Autoencoders (VAE)* wurde in [3] ein personenunabhängiges Modell für einen speziellen Wiedergaberaum gelernt, sodass anhand weniger Parameter die Impulsantworten des MIMO-Systems in diesem Raum beschrieben werden können. Die Güte des gelernten Modells hängt jedoch von der räumlichen Dichte der zum Training verfügbaren Samplingpunkte ab. Daher lässt sich schließen, dass auf Vorwissen basierende Systemidentifikationsalgorithmen nicht grundsätzlich Kalman-Filter-basierten Methoden ohne Vorwissen überlegen sind.

[1] Veröffentlichung in der Dissertation „Signal Processing Algorithms for Adaptive Loudspeaker-Based Binaural Audio Reproduction“ von Tobias Kabzinski, die sich zum Zeitpunkt dieses Berichts noch im Prozess der Begutachtung befindet.

Adaptive Crosstalk-Cancellation

Die adaptive CTC vermeidet Modellfehler, indem die CTC-Filter immer basierend auf einer Schätzung der aktuellen Lautsprecher-Ohr-Übertragungsfunktionen aktualisiert werden und somit sichergestellt wird, dass CTC-Filter und Übertragungsfunktionen zusammen passen, um den gewünschten Effekt zu erzielen. Auf Grund der algorithmischen Latenz des in [2] vorgestellten Verfahrens kann es allerdings bei sehr schnellen Kopfdrehungen dazu kommen, dass wegen der verzögerten Schätzung die auf (veralteten) Schätzungen basierend entworfenen CTC-Filtern nicht zu den momentanen Systemübertragungsfunktionen passen. Die algorithmische Latenz stellt in der betrachteten Anwendung einen kritischen Faktor für Systemidentifikationsalgorithmen dar, was die Anwendbarkeit der in [2] vorgestellten Verfahren einschränkt. Die geschätzten Zustandsraumparameter könnten, gelernt auf einer Vielzahl von verschiedenen beispielhaften Szenarien, auch für eine nahezu verzögerungsfreie Zustandsschätzung mittels Kalman-Filter in jeweils ähnlichen Szenarien angewendet werden.

Das Potential von Kalman-Filter-basierten Systemidentikationsmethoden, die geringe algorithmische Latenz aufweisen, wurde für die Anwendung in der adaptiven CTC untersucht[2]. Basierend auf trockenen HRTFs und gemessenen BRIRs (Forschungsdaten veröffentlicht in [4]) wurde simulativ untersucht, wie gut die binauralen Cues ILD und ITD reproduziert werden können. Im Unterschied zu [1] wurde dabei eine unbeschränkte Rechenleistung angenommen. Es wurden unterschiedliche Wiedergabesignale und Bewegungsgeschwindigkeiten berücksichtigt. Aus psychoakustischer Sicht ist derzeit nicht klar, inwieweit sich Ungenauigkeiten der reproduzierten binauralen Cues *während* einer schnellen Kopfdrehung auf die Plausibilität der wiedergegebenen Szene auswirken. Für moderate Bewegungen und stationäre Bedingungen liegen die reproduzierten binauralen Cues ILD und ITD nahe der Grenzen der *Just Noticable Differences (JNDs)*, sodass davon ausgegangen wird, dass dem Zuhörer das gewünschte räumliche Abbild unter diesen Umständen auch ohne Vorwissen über die BRIRs vermittelt werden kann. Insbesondere kann eine sehr schnelle Konvergenz und damit eine plausible Wiedergabe erzielt werden, sobald der Zuhörer ruht.

3.2.3 Plausible ortsstabile Wiedergabe binauraler Signale

Für die ortsstabile Wiedergabe müssen die räumlichen Eigenschaften der an den Ohren reproduzierten Signale an die Kopfbewegungen des Zuhörers angepasst werden, sodass Schallquellen unabhängig von Kopfbewegungen an einer festen Position im Raum wahrgenommen werden. Im Projekt wurde dafür eine modellbasierte, signal-adaptive Filterung der binaural aufgezeichneten Signale auf Basis von HRTF-Modellen und Headtracking-Informationen entwickelt und als Prototyp implementiert. Als Erfolgskriterium wurde eine plausible Wiedergabe angestrebt – das heißt, das binaural wiedergegebene Signal sollte einen gleichwertigen räumlichen Höreindruck zu einer realen Schallquelle, die sich mit dem Zuhörer im Raum befindet, erzeugen. Dieses Kriterium wird durch die entwickelten Verfahren in vielen Fällen erfüllt.

Die Plausibilität wurde in einem Hörversuch [5, Abschnitt 6.3] mit Design nach [6] untersucht. Zunächst wurde für jede Versuchsperson mit an den Ohren platzierten Mikrofonen eine binaurale Aufnahme angefertigt. Als Quelle für Sprachsignale wurden im Raum platzierte Lautsprecher verwendet, von denen ein oder zwei gleichzeitig aktiv waren. Im anschließenden Hörversuch trugen die Versuchspersonen Kopfhörer. Zufällig abwechselnd wurden entweder die Sprachsignale über die realen Lautsprecher oder die binaural aufgezeichneten Signale über Kopfhörer wiedergegeben. Im

[2] Diss. Tobias Kabzinski, „Signal Processing Algorithms for Adaptive Loudspeaker-Based Binaural Audio Reproduction“

Anwendungsszenario der binauralen Audio-Kommunikation repräsentiert die Lautsprecherwiedergabe die „Realität“ im Besprechungsraum. Die Kopfhörerwiedergabe entspricht der Hörerfahrung des fernen Teilnehmers. Aufgabe der Versuchspersonen war, die Kopfhörerwiedergabe anhand von Höreindrücken zu identifzieren, wobei sie den Kopf bewegen durften. Die Kopfhörerwiedergabe wurde durch die entwickelten Verfahren mittels Headtracking an die Kopfbewegungen angepasst. Bei natürlichen Kopfbewegungen fiel den Versuchspersonen die Identifikation schwer. Es kann deshalb von einer hohen Plausibilität ausgegangen werden [5, S. 127]. Dieses Ergebnis gilt für das im Versuch untersuchte Szenario mit ein oder zwei gleichzeitig aktiven Sprechern. Für das Anwendungsszenario der binauralen Audio-Kommunikation lässt sich somit schließen, dass die Qualität des binaural aufgezeichneten und über Kopfhörer ortsstabil wiedergegebenen Signals in hinreichendem Maße den Erwartungen der Versuchspersonen an eine reale akustische Szene entspricht.

Berücksichtigung von Raumeinflüssen

Den entwickelten Verfahren liegt die vereinfachende Modellannahme zugrunde, dass das binaural aufgezeichnete Signal aus gerichteten und diffusen Anteilen besteht. Da sich diese im Hinblick auf die Auswirkungen von Kopfbewegungen unterscheiden, ist es vorteilhaft, sie bei der Signalanpassung unterschiedlich zu behandeln. Für die Wahrnehmung von Schallquellen an festen Positionen im Raum sind primär die gerichteten Anteile relevant.

Eine Schwierigkeit besteht darin, dass gerichtete und diffuse Anteile in der Aufnahme i. A. als Mischung vorliegen, die nicht ohne weiteres getrennt werden kann. *Beamforming*-Ansätze nehmen diese Trennung auf Basis der unterschiedlichen räumlichen Eigenschaften (interaurale Kohärenz und Schalleinfallsrichtung) vor. Die Auswirkungen der prinzipiellen Grenzen derartiger Ansätze in der vorliegenden Problemstellung wurden theoretisch und experimentell analysiert [5, Kap. 5 u. 6]. Die Grenzen wirken sich insbesondere bei großen Kopfdrehungen des Zuhörers und stark nachhallbehafteten Aufnahmen aus und führen zu einer Verfälschung der räumlichen Eigenschaften des in der Aufnahme enthaltenen Nachhalls. Ein Verbesserungsansatz ist, zur Trennung zusätzlich zu den räumlichen Eigenschaften auch zeitliche Eigenschaften von gerichtetem Schall und Nachhall auszunutzen [7].

Hörbare Artefakte aufgrund der unzureichenden Trennung der gerichteten und diffusen Anteile können vermieden werden, indem der Phasengang des Richtungskompensationsfilters auf das ungetrennte Signal angewendet wird. Die Trennung ist dann nur noch für den Frequenzgang des Richtungskompensationsfilters wirksam [5, Abschnitt 5.6]. Interessanterweise widerspricht dieses Ergebnis der ursprünglichen Hypothese im Projektantrag, es sei vorteilhaft, „möglichst wenige Zeit-/Frequenzkomponenten überhaupt zu manipulieren, d.h. sich auf diejenigen zu fokussieren, deren Modifikation notwendig und hinreichend für die Richtungskompensation ist“.

Berücksichtigung komplexer HRTFs des Aufnahmegerätes

In den bisherigen Untersuchungen wurde für die HRTFs von Aufnahmesystem und Zuhörer eine einfache Geometrie (kugelförmiger Kopf) angenommen. Im Rahmen des Projekts wurde die Berücksichtigung komplexerer Kopfgeometrien in den entwickelten Methoden untersucht. Diese führt zu besseren Schätzungen der Modellparameter [5, Kap. 4]. Wenn die HRTFs des Aufnahmesystems nicht bekannt sind, liefert auch ein gemittelter HRTF-Datensatz bessere Ergebnisse als das Kugelmodell [8].

Anders als bei Projektbeginn angenommen ist die Berücksichtigung der HRTFs des Aufnahmegeräts keine Voraussetzung für eine plausible ortsstabile Wiedergabe. Die durch das Kugelmodell

zu erwartenden Abweichungen in den wiedergegebenen Signalen bleiben bei kleinen Kopfdrehungen (bis ca. 30°) in vielen Fällen unter der Wahrnehmungsschwelle [5, S. 70]. Entsprechend war im Hörversuch eine plausible Anpassung an kleinere Kopfdrehungen (bis ca. 30°) auch bei Nutzung des Kugelmodells möglich [5, S. 127]. Informelle Untersuchungen deuten darauf hin, dass mit allgemeinen HRTF-Modellen [8] plausible Signale auch noch mit größeren Kopfdrehungen möglich sein werden.

Aufnahme mit bewegtem Aufnahmesystem

Besteht das Aufzeichnungssystem aus Mikrofonen an den Ohren einer realen Person, ist zu erwarten, dass das Aufnahmesystem sich ebenfalls bewegt. Dies führt im aufgezeichneten Signal zu scheinbaren Bewegungen der eigentlich ortsfesten Schallquellen. Diese können durch eine angepasste signal-adaptiven Filterung kompensiert werden, wenn die Bewegungen des Aufnahmesystems durch Head-Tracking gemessen werden. In diesem Anwendungsfall ist jedoch eine robuste Vorne-Hinten-Unterscheidung wichtiger. Steht diese nicht zur Verfügung, können nur kleine Bewegungen des Aufnahmesystems ausgeglichen werden. Alternativ muss der Raumbereich, in dem sich Schallquellen befinden können, stärker eingeschränkt werden.

Zur Lösung wurden Verfahren entwickelt, mit denen durch Bewegung des Aufnahmesystems hervorgerufene Änderungen der räumlichen Signaleigenschaften zur Vorne-hinten-Unterscheidung [9] oder Elevationsschätzung [10] genutzt werden können. Für den Fall, dass die Kopfrotationsinformation nicht vorliegt, kann diese für kleine Kopfbewegungen unter Annahme einer einzelnen, feststehenden Schallquelle auch blind geschätzt werden [11]. Diese Verfahren haben ein großes Potential auch für andere Fragestellungen mit bewegten Mikrofonen (z. B. in der Signalverarbeitung für Hörgeräte).

3.3 Abweichungen und Widersprüche zu Ausgangshypothesen

Auswirkungen der Mikrofonplatzierung auf adaptive CTC bei hohen Frequenzen

Die Untersuchungen [12] und [13] legen nahe, dass die Übertragung zwischen einem ohrnahen Mikrofon, welches am Ohrkanaleingang platziert ist, und dem menschlichen Trommelfell für einen großen Frequenzbereich richtungsunabhängig ist. Diese Hypothese ist fundamental für die Funktionsweise eines adaptiven CTC-Systems mit Mikrofonen an den Ohren. Wird der Schall zwischen Mikrofonposition und Trommelfell in Abhängigkeit der Schalleinfallsrichtung unterschiedlich übertragen, ist bei Auslegung der CTC-Filter für die Mikrofonposition nur dort sichergestellt, dass die Überlagerung den gewünschten Effekt von Entzerrung und Übersprechkompensation erzielt, nicht jedoch am Trommelfell. Um die Richtungsabhängigkeit der Schallübertragung für diesen Fall zu untersuchen, wurden für ein Beispielsystem aus Neumann KU100-Kunstkopf und externem B&K 4101-B-Binauralmikrofon am Ohr Messungen der Übertragungspfade durchgeführt. Die Analyse der Richtungsabängigkeit ergab, dass im untersuchten Beispiel eine *richtungsunabhängige* Übertragung nur für Frequenzen bis 5 kHz bzw. 7 kHz am linken bzw. rechten Ohr gewährleistet war[3]. Dies könnte einen Beitrag zur permanent leicht reduzierten IC in [1] erklären. Während die o.g. Untersuchungen Abweichungen von wenigen Dezibel als nicht gravierend bewertet haben, erscheinen Abweichungen für die korrekte Funktionsweise der Übersprechkompensation hier gravierend. Diese Analyse stellt somit einen Widerspruch zu Ausgangshypothese dar. Die korrekte Lokalisierung von Schallquellen im Elevationswinkel, die auf höheren Frequenzen beruht, könnte somit eingeschränkt sein oder erfordert eine Mikrofonplatzierung näher am Trommelfell, die für eine praktische Anwendung einschränkend sein dürfte.

[3] Diss. Tobias Kabzinski, „Signal Processing Algorithms for Adaptive Loudspeaker-Based Binaural Audio Reproduction“

Headtracking-Verfahren

In Rahmen einer studentischen Abschlussarbeit[4] wurde untersucht, inwieweit das Tracking von Position und Orientierung des Zuhörers grundsätzlich auch in 3D möglich ist. Dazu wurden auf Kugelflächenfunktionen (engl. *Spherical Harmonics*, SH) basierende richtungsabhängige Laufzeitmodelle entwickelt. Die darauf basierenden Least-Squares-Probleme zur Schätzung von Position und Orientierung können als Maximum-Likelihood Schätzer interpretiert werden. Die Schätzung erfordert einige leichte Einschränkungen, z.B. die Verwendung von mindestens drei Lautsprechern, und ermöglicht keine uneingeschränkte Schätzung des Nickwinkels, da eine Änderung dieses Winkels typischerweise nicht zu einer Laufzeitänderung zum Ohr führt. Die erzielte Genauigkeit liegt bei wenigen Millimetern bzw. Grad. Daher ist keine Einschränkung gegenüber der Nutzung von traditionellen Headtrackingsystemen zu erwarten, wenn ein SH-Laufzeitmodell zum Headtracking verwendet wird.

Basierend auf den praktischen Erfahrungen mit dem PoC-System [1] gehen wir davon aus, dass eine extrem hohe Headtracking-Präzision keine praktische Systemanforderung darstellt. Die menschliche Schallquellenlokalisation bietet prinzipiell nur begrenzte Auflösung. Zudem weichen die bei der ortsstabilen Wiedergabe reproduzierten binauralen Cues wegen der Raumeinflüsse sowie Ungenauigkeit des Kugelmodells zu einem gewissen Grad von einer Referenzaufnahme ab, ohne dass dies die Plausibilität ausschließt [5, S. 127] (siehe oben). Sehr genaues Headtracking erscheint daher für eine plausible ortsstabile Wiedergabe nicht erforderlich.

Non-Uniqueness-Problem

Entgegen der Erwartung, dass das sog. *Non-Uniqueness*-Problem die Systemidentifikation grundsätzlich einschränken könnte und daher besondere Maßnahmen bei Verbesserung der Systemidentifikationsalgorithmen erfordern könnte, wurde festgestellt, dass dies in realitätsnahen Szenarien kein praktisches Problem darstellt. Wenn die binaural wiederzugebenden Signale (diffusen) Nachhall enthalten in Form von Raumimpulsantworten, die länger sind als die auf der Wiedergabeseite geschätzten Raumimpulsantworten, wirkt sich dies positiv auf die Identifizierbarkeit aus (*Tail-Effect*).

3.4 Umgang mit im Projekt entstandenen Forschungsdaten

Beispielhafte Messdaten zur Simulation einer kontinuierlichen Drehung eines Zuhörers im nachhallbehaftenen Raum wurden in [4] veröffentlicht.

3.5 Maßnahmen zur Wissenschaftskommunikation

Die im Rahmen des Projekts entwickelten Prototypen werden regelmäßig vorgeführt, zum Beispiel im Rahmen von Veranstaltungen für Schüler*innen und Studierende.

3.6 Literaturverzeichnis

[6] A. Lindau und S. Weinzierl, „Assessing the Plausibility of Virtual Acoustic Environments," *Acta Acustica united with Acustica*, Jg. 98, Nr. 5, S. 804–810, Sep. 2012. DOI: 10.3813/aaa.918562.

[12] D. Hammershøi und H. Møller, „Sound Transmission to and within the Human Ear Canal," *The Journal of the Acoustical Society of America*, Jg. 100, Nr. 1, S. 408–427, 1996. DOI: 10.1121/1.415856.

[4] Bachelorabschlussarbeit zum Thema „Modellbasiertes 3D Headtracking zur Anwendung in adaptiven Übersprechkompensationssystemen" von Henning Konermann

[13] R. Algazi, C. Avendano und D. M. Thompson, „Dependence of Subject and Measurement Position in Binaural Signal Acquisition," *Journal of the Audio Engineering Society*, Jg. 47, Nr. 11, S. 937–947, 1999.

4 Veröffentlichte Projektergebnisse

4.1 Publikationen mit wissenschaftlicher Qualitätssicherung

[1] T. Kabzinski, S. Nagel und P. Jax, „Towards a Natural Reproduction of Binaural Recordings: Combining Binaural Cue Adaptation and Adaptive Crosstalk Cancellation," in *Speech Communication; 15th ITG Conference*, (Aachen, 20.–22. Sep. 2023), VDE, Sep. 2023, S. 31–35. DOI: 10.30420/456164005.

[2] T. Kabzinski und P. Jax, „A Flexible Framework for Expectation Maximization-Based MIMO System Identification for Time-Variant Linear Acoustic Systems," *IEEE Open Journal of Signal Processing*, Jg. 5, S. 112–121, Nov. 2023, Open-Access-Publikation. DOI: 10.1109/OJSP.2023.3337721.

[3] J. Hahn, T. Kabzinski und P. Jax, „Extending Manifold-based MIMO System Identification to Adaptive Crosstalk Cancellation," in *Speech Communication; 16th ITG Conference*, (Berlin, 24.–26. Sep. 2025), zur Veröffentlichung angenommen, VDE, Sep. 2025.

[5] S. Nagel, *Interactive Reproduction of Binaurally Recorded Signals* (Aachen Series on Communication Systems 6). Düren: Shaker Verlag, Jan. 2025, ISBN: 978-3-84409-758-0.

[8] S. Nagel und P. Jax, „Evaluation of HRTF Models for Binaural Cue Adaptation," in *Speech Communication; 15th ITG Conference*, (Aachen, 20.–22. Sep. 2023), VDE, Sep. 2023, S. 166–170. DOI: 10.30420/456164032.

[9] E. Fleischhauer, S. Nagel und P. Jax, „Binaural Direction-of-Arrival Estimation Incorporating Head Movement Information," in *Proceedings of International Workshop on Acoustic Signal Enhancement (IWAENC)*, (Aalborg, Dänemark, 9.–12. Sep. 2024), IEEE, Sep. 2024. DOI: 10.1109/IWAENC61483.2024.10694619.

[10] E. Fleischhauer und P. Jax, „Full-Sphere Binaural Direction-of-Arrival Estimation Incorporating Head Rotation Information," in *Proceedings of European Signal Processing Conference (EUSIPCO)*, (Palermo, Italien, 8.–12. Sep. 2025), Open-Access-Publikation, zur Veröffentlichung angenommen, Sep. 2025.

[11] E. Fleischhauer und P. Jax, „Blind Estimation of Head Rotations From Binaural Recordings," in *Speech Communication; 16th ITG Conference*, (Berlin, 24.–26. Sep. 2025), zur Veröffentlichung angenommen, VDE, Sep. 2025.

4.2 Weitere Publikationen und öffentlich gemachte Ergebnisse

[4] T. Kabzinski und P. Jax, *Quasi-Continuous Binaural Room Impulse Responses in the Horizontal Plane: A High-Resolution Dataset*, Mai 2025. DOI: 10.5281/zenodo.15490727.

[7] E. Fleischhauer, S. Nagel, A. Balachanthiran und P. Jax, „On the Use of Dereverberation Algorithms in Binaural Cue Adaptation," in *Proceedings of German Annual Conference on Acoustics (DAGA)*, (Hannover, 18.–21. März 2024), ausgezeichnet mit dem DAGA Posterpreis, Apr. 2024, S. 1600–1603.