Quality of Experience - Evaluierung eines Telekonferenzsystems in der Entwicklungsphase

Thomas Volk, Martin Rothbucher, Klaus Diepold

Technical Report

Quality of Experience - Evaluierung eines Telekonferenzsystems in der Entwicklungsphase

Thomas Volk, Martin Rothbucher, Klaus Diepold

30. März 2014

Lehrstuhl für Datenverarbeitung
Technische Universität München

Thomas Volk, Martin Rothbucher, Klaus Diepold. *Quality of Experience - Evaluierung eines Telekonferenzsystems in der Entwicklungsphase.* Technical Report, Technische Universität München, München, 2014.

Betreut von Prof. Dr.-Ing. K. Diepold ; eingereicht am 30. März 2014 bei der Fakultät für Elektrotechnik und Informationstechnik der Technischen Universität München.

Lehrstuhl für Datenverarbeitung, Technische Universität München, 80290 München, `http://www.ldv.ei.tum.de`.

Zusammenfassung/Abstract

Deutsch

Gegenstand dieser Arbeit ist die Evaluierung der subjektiv empfundenen Qualität des Nutzers („Quality of Experience“) bei der Verwendung eines neuen, momentan am Lehrstuhl für Datenverarbeitung der TU München (LDV) entwickelten, Telekonferenzsystems. Zentrale Bestandteile dieses Systems sind die Einbindung von Lokalisierungs Quelltrennungs und Sprechererkennungs-Algorithmen bei der Aufnahme des Audiosignals, sowie eine 3-D Audiowiedergabe unter Verwendung der Binauraltechnik. Dadurch soll es dem Nutzer erleichtert werden, sich in Konferenzen mit mehreren Teilnehmern zu orientieren. Bei der Evaluierung sollen diese Bestandteile des Systems daher besonders berücksichtigt werden.

English

Scope of this thesis is the evaluation of a teleconferencing system, which is currently developed at the Institute for Data Processing, in terms of the „Quality of Experience“ percieved by the user. The system features localization, source-separation and speaker-recognition algorithms as well as an immersive audio playback to improve the users orientation ability, when participating in conferences with multiple talkers. Therefore these components of the system shall particularly be taken into consideration during the evaluation.

Inhaltsverzeichnis

Abbildungsverzeichnis

1. Einführung

Während der Entwicklung eines Telekonferenzsystems sind Experimente zur Evaluierung von wesentlicher Bedeutung, da sie wichtige Auskünfte geben, was bereits funktioniert und wo noch Schwächen des Systems liegen. In den meisten Fällen werden solche Evaluierungen in Form technischer Benchmark Tests durchgeführt. Es kann aber auch von großem Interesse sein, solch eine Evaluierung mit potentiellen Anwendern in Form subjektiver Tests durchzuführen, um einen ersten Eindruck zu bekommen, wie das System vom Anwender wahrgenommen wird.
Daher soll im Rahmen dieser Arbeit die subjektive Evaluierung eines zur Zeit am LDV entwickelten Telekonferenzsystems hinsichtlich der vom Nutzer empfundenen Quality of Experience durchgeführt werden.
Dazu soll zunächst in Kapitel 2 der Arbeit ein umfassender Überblick über das zu evaluierende System erarbeitet werden. In Kapitel 3 soll dann ein allgemeines Konzept von Qualität erarbeitet werden, bevor im Anschluss mögliche Ansätze eines Qualitätsbegriffs zur Bewertung eines Telekonferenzsystems mit binauraler Wiedergabe vorgestellt werden. Einer davon ist der für diese Arbeit zentrale Quality of Experience Begriff, der in einen Kontext zu den anderen Konzepten gestellt werden soll. In Kapitel 4 soll dann auf einige wesentliche in der Literatur behandelte Schwierigkeiten bei der Planung und Durchführung eines Experiments zur subjektiven Evaluierung mit Probanden eingegangen werden. Aufbauend auf den Kapiteln 2, 3 und 4 soll dann in Kapitel 5 das erarbeitete Konzept zur QoE Evaluierung des Telekonferenzsystems vorgestellt werden. Kapitel 6 soll dann die Ergebnisse des durchgeführten Experiments zusammenfassen, bevor in Kapitel 7 ein abschließendes Fazit gezogen und ein Ausblick auf weiterführende Arbeiten gegeben wird.

2. Systemübersicht

Zunächst soll das System, das später Gegenstand der Evaluierung sein wird, genauer vorgestellt werden. Grafik 2.1 gibt dazu einen ersten Überblick über die Verarbeitungsschritte, die das Sprachsignal von der Aufnahme bis zur räumlichen Wiedergabe durchläuft.
Von links beginnend sieht man zunächst das Mikrofon Array, das zur Aufnahme der Sprecher in der Telekonferenz dient, gefolgt von einem Audio Interface, das die analogen Signale der einzelnen Mikrofone in digitale Signale umwandelt. Diese werden dann im Block „Channel Assignement“, auf den unter Punkt 2.1 genauer eingegangen wird, verarbeitet, sodass jeder Teilnehmer der Konferenz auf einem separaten Kanal übertragen werden kann. Dann folgt die Übertragung über das Netzwerk, die im Rahmen dieser Arbeit mangels einer netzwerkfähigen Implementierung nicht weiter betrachtet werden soll. Nach der Übertragung folgt dann auf der Empfängerseite die Verarbeitung der Audio Streams durch die unter 2.2 auf Seite 13 beschriebene 3-D Audio Engine und schließlich die Rückwandlung des digitalen Signals in ein analoges zur räumlich aufbereiteten Wiedergabe auf den ganz rechts eingezeichneten Kopfhörern.

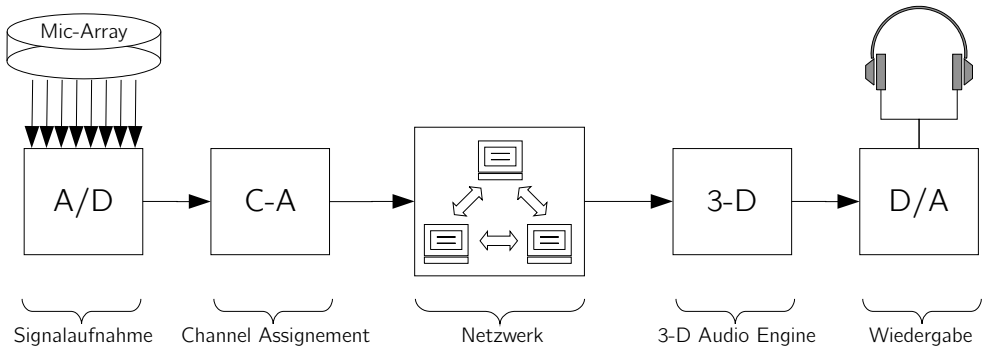


Abbildung 2.1.: Schematische Übersicht der Audio-Signalverarbeitung von der Aufnahme (links) bis zur Wiedergabe auf dem Kopfhörer (rechts)

2.1. Channel Assignement

Grundvoraussetzung für eine räumliche Wiedergabe der einzelnen Konferenzteilnehmer ist, dass die Sprachinformation jedes einzelnen Teilnehmers in einem separaten Audio Stream verfügbar ist. Es soll also jedem Sprecher durch das System ein eigener Kanal zu-

gewiesen werden, weshalb im Rahmen dieser Arbeit entsprechend [48] der Begriff „Channel Assignement“ verwendet wird.

2.1.1. Headset Szenario

Der einfachste Fall für das Channel Assignement soll in dieser Arbeit als „Headset Szenario“ bezeichnet werden. Dabei hat jeder Sprecher - wie bei der Verwendung eines Headsets üblich - ein eigenes Mikrofon. Die Sprachinformation eines einzelnen Sprechers liegt also bereits separat auf einem eigenen Kanal vor.
Dieses Szenario ist besonders dann realistisch, wenn die Sprecher sich nicht am gleichen Ort befinden. Sollten sich aber mehrere Konferenzteilnehmer am gleichen Ort aufhalten, so kann es sinnvoll sein, dass diese gemeinsam ein Mikrofon oder ein Mikrofon Array („Konferenzspinne“) benutzen.

2.1.2. Konferenztisch Szenario (Algorithmus 4)

Die Verwendung eines Mikrofon Arrays bietet den Vorteil, dass die Anzahl der Teilnehmer flexibel bleibt und nicht durch die verfügbaren Mikrofone begrenzt ist. Zudem gestaltet sich die Interaktion zwischen den Teilnehmern, wenn diese z.B. um einen Konferenztisch versammelt sind, auf diese Weise natürlicher.
Allerdings stellt sich bei Verwendung einer Konferenzspinne die Frage, wie man aus dem aufgenommenen Signal die getrennten Audio Streams der einzelnen Sprecher erhält. Im Rahmen verschiedener Projekte wurden dazu am LDV unterschiedliche Ansätze erforscht, implementiert und evaluiert. Eine systematische Übersicht zu diesen Arbeiten findet sich in [41]. In [27] und [48] wurden diese Ansätze dann nochmals überarbeitet und zu einer zusammenhängenden MATLAB-Implementierung vereint.
Im Folgenden soll eine kurzer Überblick über diese Implementierung gegeben werden. Vorab sei noch vermerkt, dass in [27] jeweils 2 Verfahren zur Lokalisierung und zur Quelltrennung untersucht wurden und ein „Blind Source Separation“ Algorithmus, der beide Aufgaben zugleich löst. Im Rahmen dieser Arbeit soll jedoch nur auf eine Kombination der Verfahren eingegangen werden, die sich bei der Evaluierung in [27] anhand von SIR, SAR und SDR Werten als am besten geeignet herausstellte.
Am Eingang liegen die digitalisierten Audiosignale der 8 Mikrophone der Konferenzspinne (vgl. Abbildung 2.1 auf der vorherigen Seite). Aus diesen wird zunächst mit Hilfe des „SRP-PHAT“ (Steered Response Power Phase Transform) Algorithmus die Positionsinformation berechnet. Die Trennung eventuell gleichzeitig sprechender Teilnehmer wird dann von der „Geometric Source Separation“(GSS) vorgenommen. Wenn sich die Sprecher während der Konferenz nicht bewegen, kann so bereits das Channel Assignement erfolgen. Sollte jedoch ein Sprecher während der Konferenz die Position wechseln, würde er einem falschen Kanal zugeordnet. Deshalb ist anschließend noch die Sprechererkennung implementiert. Diese überprüft, ob der Sprecher auf dem ihm zugewiesenen Kanal noch mit seinen gespeicherten Merkmalen („Mel-Frequency Cepstral Coefficients“, MFCC) überein-

stimmt. Wenn dies nicht mehr der Fall ist, kann eine Anpassung des Channel Assignement vorgenommen werden.
Neben der Variante ohne und der mit Sprecher Erkennung wurden in [48] noch 4 weitere Algorithmen verglichen. Der eben beschriebene Algorithmus mit SRP-PHAT, GSS und der Sprechererkennung lieferte jedoch die besten „Diarization Error"-Werte [48]. Und auch erste informelle subjektive Tests mit den verschiedenen Algorithmen bestätigten diese Tendenz eindeutig. Deshalb wurde dieser Algorithmus zum Channel Assignement für die weitere Evaluierung gewählt.

2.2. 3-D Audio-Display

Nachdem nun idealerweise jeder Sprecher auf einem eigenen Kanal übermittelt werden kann, sollen im folgenden Abschnitt die Elemente des Systems betrachtet werden, die es ermöglichen, dem Anwender die zugeschalteten Konferenzteilnehmer mit Hilfe der „Binauraltechnik" räumlich getrennt wiederzugeben. Unter Binauraler Wiedergabe (oft auch „Immersive Audio") versteht man, wie in [14] beschrieben, das simulieren eines räumlichen Schallfeldes durch die Faltung monophoner Quellen mit Übertragungsfunktionen, welche die richtungsabhängigen Reflexionen des äußeren Ohres und des Torso für linkes und rechtes Ohr nachbilden. Anhand dieser Reflexionen und der zeitlichen Verzögerung zwischen linkem und rechtem Ohr sowie des Pegelunterschiedes ist der Mensch in der Lage, den Ursprungsort einer Schallquelle zu bestimmen. Die Übertragungsfunktionen nennt man „Head Related Transfer Functions" (HRTFs).
Die wesentlichen Systemkomponenten für die binaurale Wiedergabe sind zum einen das „Head-Tracking"zum Erfassen der Kopfbewegungen des Anwenders und zum anderen die „3D-Audio-Engine", die die Faltung der übermittelten Sprecher-Streams mit den Übertragungsfunktionen aus einer HRTF-Datenbank durchführt. Im Abschnitt 2.2.3 auf Seite 15 soll dann genauer auf die Gewinnung der HRTF Datenbank sowie verschiedene Möglichkeiten für eine individuelle Anpassung der HRTFs an den Nutzer eingegangen werden. Abschließend soll dann noch kurz die am LDV aufgenommene LDV HRTF Database vorgestellt werden.

2.2.1. Head-Tracking

Das Head-Tracking soll es ermöglichen, dass die Schallquellen auch dann, wenn der Zuhörer seinen Kopf bewegt, ihre Position im virtuellen Schallfeld beibehalten. Es soll also verhindert werden, dass sie sich mit dem Kopf der Person mitbewegen. Damit soll einem Konferenzteilnehmer ein realistischeres räumliches Hörerlebnis ermöglicht werden.
Besonders hervorgehoben sei an dieser Stelle der positive Effekt des Head-Trackings hinsichtlich der Lokalisationsgenauigkeit. Studien hierzu finden sich z.B. in [45], [50], [11] und [35]. Im Rahmen von Hörversuchen wurde in all diesen Arbeiten eine Verbesserung der Lokalisation durch das Verwenden von Head-Tracking festgestellt. Insbesondere konnte

ein deutlicher Rückgang der sogenannten „Front-Back Confusions“- also des fälschlichen Lokalisierens einer Quelle in der vorderen statt der hinteren Hemisphäre oder umgekehrt - verzeichnet werden.
Für ein Telekonferenzsystem, dass dem Anwender mit Hilfe eines durch die Binauraltechnik erzeugten dreidimensionalen Schallfeldes unter anderem eine bessere Orientierung ermöglichen soll, erscheint die Verwendung eines Tracking Systems also überaus sinnvoll.

Head-Tracking mit einer monokularen USB-Kamera

Die erste Implementierung zum Head-Tracking für das betrachtete System soll die Kopfposition des Teilnehmers aus dem zweidimensionalen Bild einer handelsüblichen monokularen USB-Kamera („Webcam“) ermitteln. Dieser Ansatz bietet den großen Vorteil, dass er auf allen heutzutage gängigen Rechnern preiswert und einfach umgesetzt werden kann. Eine genaue Beschreibung der Implementierung findet sich in [18]. Kurz zusammengefasst wird dabei zunächst versucht auf dem Bild ein Gesicht zu erkennen und mit Hilfe eines gegebenen Feature Sets darin Augen, Nase und Mund zu markieren. Daraus kann bereits eine Schätzung der Kopfposition ermittelt werden, die dann zur Initialisierung der „Head Pose Estimation“ [18] verwendet werden kann. Diese soll dann mit einer größeren Zahl an Gesichts Features die genaue Position ermitteln.
In der Vorbereitungsphase der Versuche (s. Kapitel 5 auf Seite 33) stellte sich jedoch heraus, dass - bei verschiedenen Personen und abhängig von der Beleuchtung - regelmäßig links/rechts Verwechslungen beim Erfassen der Kopfbewegung auftreten. Weiterhin schlug die Reinitialisierung des Systems nach Verlust der Featurepunkte im Bild häufig fehl. Daher stand für die Versuche lediglich die oben beschriebene Initialisierungsroutine, welche eine erste Schätzung der Kopfposition anhand von Augen, Nase und Mund liefert, zur Verfügung. Diese lieferte weitaus zuverlässigere Daten hinsichtlich der Orientierung und war stets in der Lage, sich bei Verlust der Featurepunkte zu reinitialisieren. Diese zwei Voraussetzungen wurden als unverzichtbar für die Durchführung der geplanten Hörversuche angesehen.
Da dieses Trackingkonzept darauf basiert, das Gesicht einer Person auf dem Bild der USB Kamera zu erkennen ist der Bewegungsradius natürlich begrenzt. Damit sich das Gesicht im Blickfeld der Kamera befindet sollte ein Teilnehmer nicht weiter als $\pm 35^\circ$ nach links bzw. rechts schauen. Daher wurde das Tracking der Bewegung auf diesen Wertebereich beschränkt. Hinsichtlich auf- und abwärts Bewegungen des Kopfes wurde die mögliche Position auf die maximalen Elevationswinkel der LDV HRTF Database (s. auch Abschnitt 2.2.4 auf Seite 17), also auf -10° und $+40^\circ$, begrenzt. Auf eine Bestimmung der Neigung des Kopfes wurde verzichtet, da die Werte für den Roll-Winkel sich als auffällig instabil erwiesen. Wenn der gegebene Wertebereich verlassen wird, wird die Position auf den Grenzwert gesetzt und gewartet, bis wieder Positionsdaten innerhalb des erlaubten Radius auftreten.
Laut [18] arbeitet die Bildverarbeitung dieses Systems mit $23 fps$.

Head-Tracking mit einem Infrarot Kamera-System

Das zweite System, das am LDV für das Head-Tracking zur Verfügung steht, ist das „DTrack2“ der Firma „AR Tracking“[1] . Es arbeitet mit 3 Infrarotkameras, die die Position eines „Tracking Bodys“- also eines kalibrierten Gegenstandes, der mit Infrarot reflektierenden Kugeln bestückt ist - erfassen. Das DTrack2 kann sowohl die Position einer einzelnen Kugel im Raum bestimmen (X, Y, Z, 3 Freiheitsgrade), als auch Position und Orientierung (X, Y, Z und φ, θ, γ, 6 Freiheitsgrade) eines Tracking Bodys. Die Infrarot Kameras liefern dabei maximal 60*fps* [7].
Bringt man nun einen solchen Tracking Body am Kopf einer Person an, kann man deren Position und Orientierung bestimmen, solange sie sich im Blickfeld von mindestens zwei Kameras aufhält. Der Vorteil des DTrack2 besteht darin, dass es ein lückenloses, nahezu fehlerfreies Tracking liefert und dem Nutzer einen weitaus größeren Aktionsradius als die USB-Kamera basierte Variante erlaubt. Die Kameras können bis zu 4*m* vom Tracking Body entfernt sein [7]. Für ein reales Telekonferenzsystem kommt das DTrack2 aufgrund des hohen technischen Aufwands und vor allem wegen der unverhältnismäßig hohen Anschaffungskosten jedoch nicht in Frage. Es kann im Laufe der Entwicklung eines solchen Systems aber durchaus zu Vergleichszwecken verwendet werden.

2.2.2. Convolution Engine

Zur Faltung der Sprecher Streams mit den HRTFs steht am LDV eine C Implementierung der Convolution Engine zur Verfügung, die im Rahmen des SFB 453 Projekts[2] der Deutschen Forschungsgesellschaft (DFG) zum Thema „Wirklichkeitsnahe Telepräsenz und Teleaktion“ entwickelt wurde. Diese ermöglicht es unter Berücksichtigung der Positionsdaten eines Tracking Systems (s. Abschnitt 2.2.1 auf Seite 13) eine oder mehrere Quellen in Echtzeit zu verarbeiten und binaural aufbereitet wiederzugeben.

2.2.3. HRTF Datenbank

Nachdem nun das Head-Tracking und die Implementierung des Faltungsalgorithmus beschrieben wurden, sollen im letzten Abschnitt zum 3D Audio-Display verschiedene Möglichkeiten eine HRTF-Datenbank zu erstellen und individuell an den Nutzer anzupassen zusammengefasst werden.

Kunstkopf HRTFs

Die erste hier vorgestellte Variante, HRTF-Daten zu ermitteln, ist die Verwendung eines Kunstkopfes bei den Aufnahmen. Ein Anwendungsbeispiel findet sich in [25]. Und auch in [9] wurden die HRTFs eines Kunstkopfes aufgenommen. Der wohl gängigste - und auch

[1] http://www.ar-tracking.com/products/tracking-systems/arttrack-system (17.07.2013)
[2] www.sfb453.de (17.07.2013)

in [25] und [9] verwendete - Kunstkopf ist der von der Firma „G.R.A.S. Sound & Vibration“ vertriebene KEMAR [3].
Der KEMAR bietet gegenüber einer Versuchsperson den Vorteil, dass man ihn ohne weiteres langwierigen Aufnahmeprozeduren unterziehen kann und er sich während den Aufnahmen nicht bewegt oder Geräusche von sich gibt. Der wesentliche Nachteil besteht allerdings darin, dass die auf diesem Wege erhaltenen HRTFs in keiner Weise individuell angepasst sind. Die HRTFs verschiedener Personen weisen jedoch in der Regel wesentliche spektrale Unterschiede auf [14], die bei diesem Ansatz gänzlich vernachlässigt werden.

HRTF-Auswahl durch den Nutzer (DOMISO)

Eine erste Möglichkeit eine individuelle Anpassung an die jeweilige Person vorzunehmen ist ein subjektives Auswahlverfahren („HRTF Selection“). Dabei soll der der Teilnehmer im Rahmen eines Hörversuchs aus einer Datenbank denjenigen HRTF-Datensatz auswählen, der am besten zu ihm passt. Voraussetzung hierfür ist natürlich, dass bereits eine solche Datenbank mit mehreren Datensätzen verfügbar ist. Zwei Varianten solcher Auswahlverfahren sind in [46] und [29] beschrieben.
Am LDV wurde das Verfahren aus [29] in leicht abgewandelter Form implementiert [26]. Dabei bekommt die Versuchsperson in paarweisen Vergleichen Testgeräusche, die mit den unterschiedlichen HRTF Datensätzen bearbeitet wurden vorgespielt und soll sich jeweils für einen der beiden entscheiden. Der Versuch ist als „Swiss Style Tournament“ [29] angelegt, sodass mit möglichst geringem zeitlichen Aufwand aus 12 Datensätzen der am besten geeignete ermittelt werden kann.

Anpassung der HRTFs anhand anthropometrischer Daten durch Regression

Eine weitere Möglichkeit, aus einer HRTF Datenbank einen individuell an den Nutzer angepassten Datensatz zu gewinnen, ist die Berechnung neuer Übertragungsfunktionen aus einer vorhandenen Datenbank durch ein auf anthropometrische Daten gestütztes Regressionsverfahren. In [31] wurde am LDV ein solches Verfahren implementiert, mit dem anhand von 8 anthropometrischen Werten mittels „Partial Least Squares Regression“ (PLSR) ein neuer HRTF Datensatz für eine Person berechnet werden kann.
Das Verfahren bietet den Vorteil, dass die anthropometrischen Daten sehr schnell messbar sind und dem Nutzer das Aufnahmeprozedere in einem Labor erspart bleibt. Grundvoraussetzung ist auch hier eine verfügbare HRTF Datenbank, die den Untersuchungen in [31] zufolge mindestens 15 Datensätze enthalten sollte.

[3] www.kemar.us (17.07.2013)

Verfahren zur Gewinnung der HRTFs	Anzahl der benötigten Datensätze	Zeitaufwand für den Nutzer	Individuelle Anpassung
KEMAR	1	-	nein
DOMISO	12	ca. 20 min	ja
PLSR	>15	ca. 5 min	ja
Individuelle IMessung	1	ca. 50 min	optimal

Tabelle 2.1.: Vergleich der Verfahren zur Gewinnung von HRTFs für die binaurale Wiedergabe

Individuelle HRTFs

Die vierte und letzte in dieser Arbeit betrachtete Möglichkeit einen HRTF Datensatz für einen Konferenzteilnehmer zu ermitteln ist schließlich das Aufnehmen der eigenen HRTFs der Person. Der offensichtliche Vorteil dieses Ansatzes besteht darin, dass so die maximale individuelle Anpassung der HRTFs gegeben ist. Allerdings steht diesem Vorteil ein deutlich gesteigerter zeitlicher Aufwand für jeden Teilnehmer gegenüber.

2.2.4. Die LDV HRTF Database

Alle in dieser Arbeit verwendeten HRTFs sind in der LDV HRTF Database enthalten, oder via PLSR aus ihr errechnet worden. Deshalb soll die Datenbank hier kurz vorgestellt werden.
Die HRTFs wurden im Halbfreifeldraum des LDV aufgenommen. Eine genaue Beschreibung der Aufnahmebedingungen findet sich in [49]. Insgesamt wurden 35 Personen und 1 KEMAR Kunstkopf vermessen. Dabei wurden jeweils 6 Elevationsebenen $(-10^\circ, 0^\circ, 10^\circ, 20^\circ, 30^\circ, 40^\circ)$ entsprechend dem in [42] beschriebenen Verfahren kontinuierlich aufgenommen sowie die für die Regression (s. Abschnitt 2.2.3 auf der vorherigen Seite) benötigten anthropometrischen Daten der Versuchspersonen gemessen.
Durch die Aufnahme eines kontinuierlichen Anregungssignals bei Rotation der Versuchsperson auf einem Drehteller lassen sich für jede der Elevationsebenen HRTFs in nahezu beliebiger Auflösung berechnen. Entsprechend [23] und weiterer Untersuchungen z.B. in [33] wurde eine horizontale Auflösung von 1° gewählt, da diese kleiner als die für die meisten Menschen wahrnehmbare Mindestschrittweite ist. Die vertikale Auflösung von 10° liegt zwar oberhalb des in [33] empfohlenen Abstands von 4°, stellt aber einen guten Kompromiss zwischen zeitlich realisierbarem Aufwand für die Versuchspersonen und dem gewünschten Aktionsradius für ein Telekonferenzsystem dar.
Erwähnt sei an dieser Stelle auch noch, dass die HRTFs - wie z.B. auch in [25] - entzerrt wurden, um den Einfluss der Übertragungsstrecke (Lautsprecher, Raum und Mikrofon) und des Kopfhörers bei der Wiedergabe auszugleichen. Dazu wurden jeweils eine Aufnahme mit einem freistehenden Mikrophon, an der Position, an der sich sonst der Mittelpunkt der Ohr zu Ohr Linie einer Versuchsperson befand, und eine Aufnahme mit Kopfhörern und

den Mikrophonen am Ohr des Probanden gemacht. Die aus diesen Aufnahmen errechneten Impulsantworten wurden anschließend nach einem in [20] vorgestellten Verfahren für die Entzerrung invertiert.

2.3. Zusammenfassung

Nachdem die Komponenten des Systems nun einzeln vorgestellt wurden sollen noch einmal kurz die wesentlichen Funktionen hervorgehoben werden.
Das System bietet also die Möglichkeit mehrere Sprecher, die von einer eigens angefertigten Konferenzspinne aufgenommen wurden, durch einen Channel Assignement Algorithmus zu lokalisieren, zu trennen und jeweils einem eigenen Kanal zuzuweisen. Die daraus einzelnen Sprecher Streams werden dann auf der Wiedergabeseite mit Hilfe der Binauraltechnik räumlich getrennt abgespielt, wobei der Nutzer auf vier verschiedene Optionen bei der Auswahl HRTF Datenbank zurückgreifen kann. Zudem besteht die Möglichkeit, die Simulation des räumlichen Schallfeldes mit einem Head-Tracking System zu unterstützen. Eine abschließende Übersicht findet sich in Abbildung 2.2.

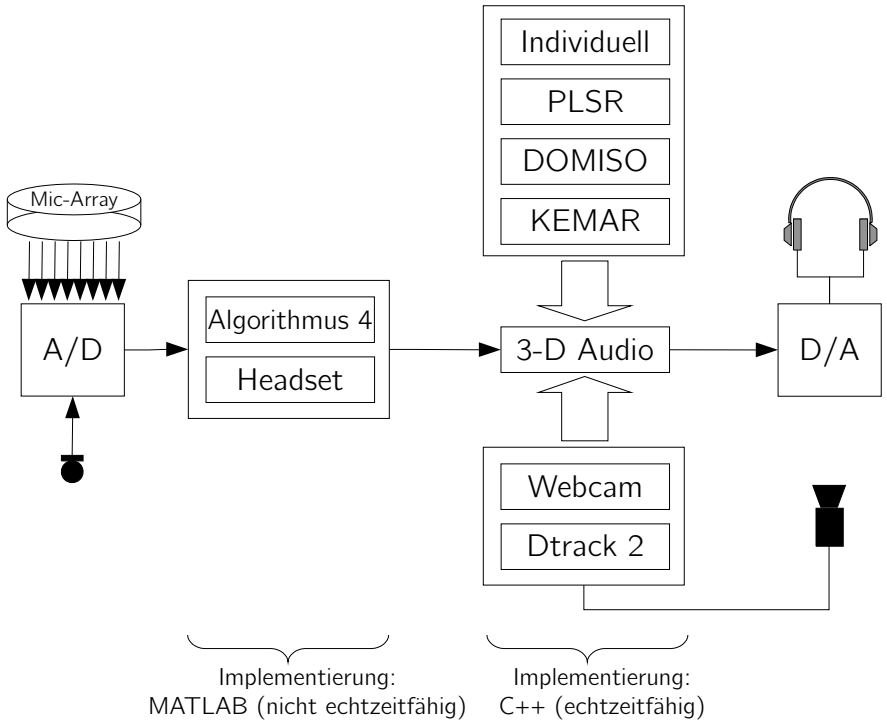


Abbildung 2.2.: Zusammenfassende Übersicht über das Telekonfernezsystem mit den vorgestellten Varianten zum Channel Assignement, Head-Tracking und für die binaurale Wiedergabe

3. Qualität und grundsätzliche Evaluierungsansätze

Um später ein Konzept zur Evaluierung der Qualität des Systems erarbeiten zu können sollen in diesem Abschnitt der Arbeit einige Ansätze zur Definition eines Qualitätsbegriffs betrachtet werden. Dabei sollen zunächst einige Überlegungen zum Begriff Qualität im Allgemeinen zusammengefasst werden. Im zweiten Abschnitt soll dann die Qualität im Bezug auf Telekonferenzsysteme betrachtet werden. Da die Audiowiedergabe einen wesentlichen Teil des betrachteten Systems ausmacht sollen dann im dritten Abschnitt mögliche Kriterien zur Definition von Klangqualität - sowohl allgemein als auch im Bezug auf die Binauraltechnik - vorgestellt werden, bevor abschließend ein Überblick zu Inhalt und Entstehung des noch relativ jungen Begriffes „Quality of Experience“ (QoE) gegeben werden soll.

3.1. Definitionen eines allgemeinen Qualitätsbegriffes

Die Frage was die Qualität einer Sache eigentlich ausmacht und wie sie sich messen und vergleichen lässt, wird schon seit geraumer Zeit wissenschaftlich bearbeitet. Die ersten Ansätze dazu finden sich in den 1960er Jahren auf dem Feld der Nahrungsmittelforschung. In Kapitel 2.4 in [36] findet sich ein guter Überblick über verschiedene Aspekte des Begriffs Qualität und deren Verknüpfungspunkte, die hier kurz wiedergegeben werden sollen. Als Ausgangspunkt dienen darin folgende vier Definitionen verschiedener Aspekte des Begriffs Qualität:

Qualität als Qualitas d.h. die sich aus ihre charakteristischen Elgenschaften ergebende Beschaffenheit einer Sache. Solche Eigenschaften sind z.B. das Material, der Preis, die Robustheit oder die Farbe.

Qualität als Güte also als Ausdruck, für das intuitives Empfinden des Menschen, wie gut eine Sache ist.

Qualität als Standard d.h. inwieweit erfüllt eine Sache vorab definierte Anforderungen und/oder damit verbundene Bedürfnisse des Menschen.

Qualität als Ereignis Damit ist hier der subjektive Eindruck des Menschen beim Umgang mit einer Sache gemeint.

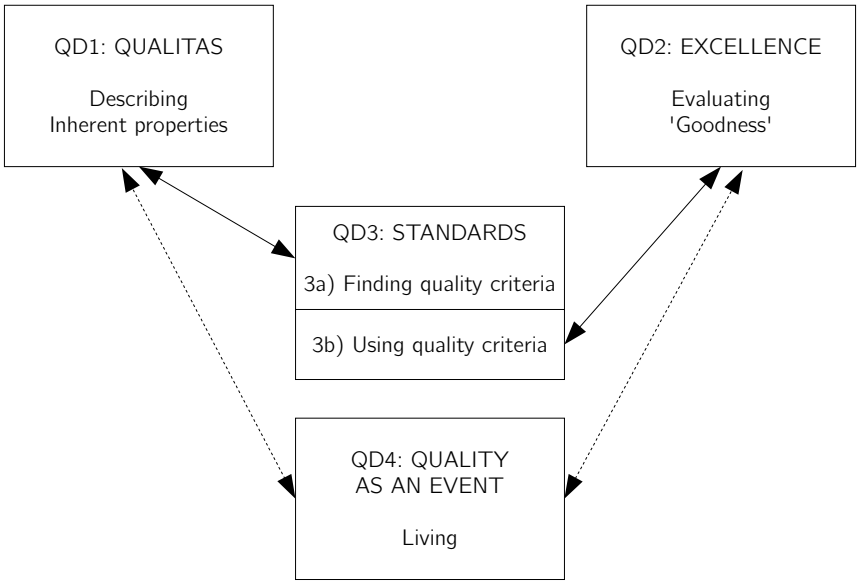


Abbildung 3.1.: Die vier Qualitäts-Definitionen aus [36] und ihre Verknüpfung

Die erste Definition hat dabei objektive Eigenschaften einer Sache im Auge, während die zweite die subjektive Wahrnehmung des Menschen hinsichtlich einer Sache in den Vordergrund stellt. Die dritte Definition der Qualität als Standard versucht eine Verbindung zwischen den beiden herzustellen. Diese Verbindung erfolgt, indem versucht wird, Kriterien für die Qualitas einer Sache zu finden, welche die subjektive Wahrnehmung des Menschen hinsichtlich der Güte der Sache wiedergeben. Die vierte Definition schließlich definiert Qualität als Ereignis, das vom Menschen erlebt wird. Der Mensch schlüpft also in die Rolle eines Konsumenten bzw. Anwenders einer Sache, die bestimmte Eigenschaften im Sinne der Qualitas besitzt und der er selbst einen gewissen Grad an Güte zuordnen kann. Diese können zwar einen Einfluss auf die subjektiv erlebte Qualität haben, reichen aber nicht aus, um diese umfassend zu beschreiben. Die aus [36] entnommene Grafik in Abbildung 3.1 soll die Zusamenhänge zwischen den vier Definitionen nochmals verdeutlichen.
Aufbauend auf diesem allgemeinen Konzept soll nun im Rest des Kapitels konkreter auf den Anwendungsfall der Qualitäts Evaluierung des Telekonferenzsystems im Rahmen der vorliegenden Arbeit eingegangen werden.

3.2. Klangqualität

Der subjektive Eindruck eines Teilnehmers bei einer Telekonferenz basiert in erster Linie auf dem wahrgenommenen Klangereignis, zu dem die einzelnen Komponenten des

Systems beitragen. Daher soll in diesem Abschnitt ein kurzer Überblick über mögliche Definitionen des Begriffs Klangqualität gegeben werden. (Viele Telekonferenzsysteme bieten auch eine visuelle Komponente, z.B. per Webcam. Diese Option bietet das vorliegenden Systems jedoch nicht. Daher soll sie im Rahmen der vorliegenden Arbeit nicht weiter berücksichtigt werden.)
Ein erster umfassender Ansatz, die Qualität von Klangereignissen zu ermitteln findet sich in [32]. Der Autor (T. Letowski) gelangt darin zu folgender Definition von Klangqualität:

> „*Sound quality is that assessment of auditory image in terms of which the listener can express satisfaction or dissatisfaction with that image. Sound quality can be judged by comparing images produced by several external stimuli or by referencing a percieved image to the concept residing in the listeners memory.*“

Dabei nimmt er bereits einige wesentliche Elemente vorweg, die sich auch unter Punkt 3.5 auf Seite 24 in der Definition des Quality of Experience Begriffes wiederfinden, nämlich die subjektive Bewertung des Hörers („*satisfaction*“) sowie den Vergleich als wesentlichen Bestandteil der Bewertung. Des Weiteren werden in [32] zwei Ansätze zur Bewertung vorgeschlagen: Ein „globaler“ und ein „parametrischer“. Besonders auf letzteren soll hier verwiesen werden, da er sich auch in zahlreichen späteren Veröffentlichungen zur Bewertung von Klangqualität - wie beispielsweise in [13] und [43] - wiederfindet. Abbildung 3.2 auf der nächsten Seite zeigt das „*Mural*“ aus [32] um den parametrischen Bewertungsansatz zu veranschaulichen. Durch die Zerlegung eines Klangereignisses in mehrere Parameter soll versucht werden, dem multidimensionalen Charakter von Klangereignissen gerecht zu werden. Eine wesentliche Schwierigkeit besteht jedoch darin, bei der parametrischen Bewertung ein Einverständnis zwischen dem Entwickler des Experiments und den Teilnehmern über die Bedeutung der einzelnen Parameter zu erreichen. Daher wurden Prozeduren entwickelt, um mit einem Panel an Versuchsteilnehmern ein geeignetes Vokabular zur Bewertung zu erarbeiten. Die Teilnehmer bei einem solchen Panel sind in der Regel erfahrene Hörer oder haben zumindest ein ausführliches Training hinsichtlich der in dem Hörversuch zu bewertenden Phänomene absolviert. Die gesteigerte Präzision des Vokabulars erkauft man sich also mit einem erheblichen zeitlichen und damit auch finanziellen Aufwand. Ein aktueller Überblick zu solchen Verfahren sowie ein neu entwickeltes Verfahren und Anwendungsbeispiele für den Bereich der Binauraltechnik finden sich in [34].
Neben der direkten Befragung von Versuchspersonen gibt es auch noch einen weiteren Weg zur Bestimmung der Klangqualität. Dabei liefern psychoakustische Experimente die Grundlage zur Modellierung von Kenngrößen, die dann bei der Anwendung aus physikalischen Messungen berechnet werden können. Ein Überblick über solche Verfahren und ihre Anwendung findet sich in [22]. Sehr häufig finden solche Verfahren Anwendung im Bereich des Lärmschutz oder beispielsweise auch beim akustischen Design von Kraftfahrzeugen. Für die Bewertung räumlicher oder insbesondere simuliert räumlicher Klanger-

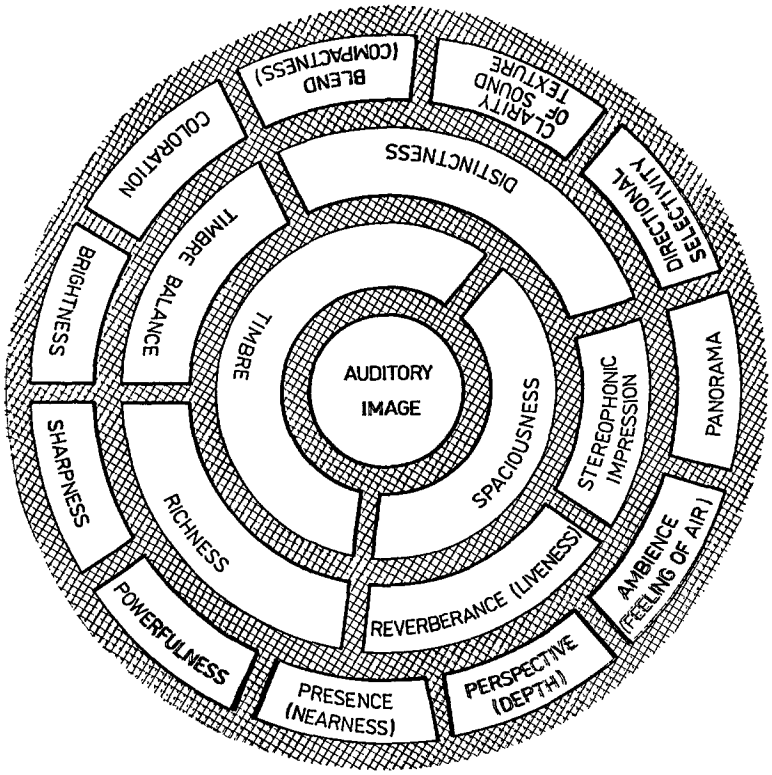


Abbildung 3.2.: Das „*Mural*“ aus [32] soll die von Letowski vorgeschlagene Aufteilung des Klangbildes in parametrische Eigenschaften verdeutlichen

eignisse fanden sich bei der Recherche zur vorliegenden Arbeit jedoch keine geeigneten Evaluierungsansätze dieser Art.

3.3. Qualität von HRTFs

Da die Binauraltechnik bei dem betrachteten Telekonferenzsystem einen wesentlichen Beitrag zur Klangwahrnehmung des Nutzers leistet, soll an dieser Stelle auch ein Überblick über gängige Metriken zur Evaluierung der Qualität von HRTFs gegeben werden.
Phänomene wie der Cocktailparty Effekt [14] oder allgemein die Fähigkeit des Menschen Schallquellen mit seinem Gehör zu lokalisieren, sind der Grund dafür, dass die Forschung sich mit der Binauraltechnik auseinandersetzt. Da die charakteristischen Merkmale der HRTFs abhängig von der Position der Schallquelle in Bezug auf den Kopf sind, ist es naheliegend zu überprüfen wie originalgetreu die Richtungswahrnehmung eines Probanden bei der binauralen Wiedergabe ist. Daher sind eines der am weitesten verbreiteten Verfahren zur Evaluierung von HRTFs sogenannte Lokalisationstests. Dabei wird einer

Person ein binaural aufbereiteter Stimulus vorgespielt, dessen Position der Proband dann anschließend zuordnen soll. Durch mehrfache Wiederholung des Experiments mit unterschiedlichen Quellpositionen kann dann aus der Abweichung zwischen den tatsächlichen Positionen und der jeweiligen subjektiven Schätzung ein Lokalisierungsfehler berechnet werden. Anwendungsbeispiele finden sich z.B. in [11] und [28].
Ein anderer Ansatz ist die anwendungsbezogene Evaluierung der HRTFs, wenn sie beispielsweise für die räumliche Wiedergabe in der Telekommunikation eingesetzt werden. Hier gibt es Studien, die die Sprecher Identifizierung („wer spricht?“) [15] oder die Verständlichkeit („was wird gesprochen?“) [17], [21] untersuchen. Außerdem gibt es Untersuchungen, bei denen die Teilnehmer Aufgaben mit der Unterstützung eines 3-D Audio Systems bewältigen sollen. Die Performance bei der Lösung der Aufgaben unter Verwendung verschiedener Wiedergabeverfahren liefert dann das Qualitätsurteil. Anwendungsbeispiele findet sich in [30] und [17].

3.4. Quality of Service für Telekonferenzsysteme

Der derzeit gängigste Ansatz zur Evaluierung von Telekonferenzsystemen stellt die Frage nach der „Quality of Service“ (QoS). Der aktuelle Standard der ITU dazu ist die ITU-T E.800 [5]. Darin wird die QoS festgelegt als:

> „*(...) totality of characteristics of a telecommunications service that bear on its ability to satisfy stated and implied needs of the user of the service.*“

Diese Definition verfolgt weitestgehend das Konzept der „Qualität als Standard“, wie in Abschnitt 3.1 auf Seite 19 vorgestellt. Bei der Bewertung der QoS hält sich die ITU dabei in erster Linie an Kriterien, die auf die Performance bezogen sind, wie zum Beispiel Geschwindigkeit oder Fehlerwahrscheinlichkeiten. Da das hier betrachtete System aber noch nicht in einer netzwerktauglichen Implementierung vorliegt, ist eine Evaluierung anhand dieser Kriterien nicht sinnvoll.
Als Teilbereich der QoS definiert [5] neben den eben erwähnten Aspekten [5] des Weiteren den Begriff der „Quality of Service experienced/percieved by customer/user“ (QoSE) wie folgt:

> „*A statement expressing the level of quality that customers/users believe they have experienced.*
> *NOTE 1 – The level of QoS experienced and/or perceived by the customer/user may be expressed by an opinion rating.*“

Neben den objektiven, auf die Performance bezogenen Kriterien wird hier also auch die Notwendigkeit einer subjektiven Bewertung durch den Nutzer gesehen. Als subjektive

Kriterien werden im Rahmen der ITU-T E.800 im wesentlichen die Sprachqualität entsprechend [6] sowie der „Mean Opinion Score“ (MOS) entsprechend [1] empfohlen.
Mit dem QoSE Teilaspekt geht die ITU bereits einen ersten Schritt in Richtung des Quality of Experience Konzepts, das im folgenden Abschnitt dieser Arbeit genauer vorgestellt werden soll.

3.5. Quality of Experience

Ein relativ junger Ansatz zur Beurteilung von Qualität ist die Quality of Experience (QoE), der eine ähnliche Idee wie die der „Qualität als Ereignis“ (wie unter Punkt 3.1 auf Seite 19 beschrieben) zugrunde liegt. Die Definition des QoE Begriffs im Rahmen dieser Arbeit orientiert sich weitestgehend am „Qualinet White Paper on Definitions of Quality of Experience and Related Concepts“ [19] aus dem Jahr 2012.
QoE kann demnach als eine Ergänzung der in Punkt 3.4 auf der vorherigen Seite bereits erwähnten QoS bei der Suche nach einem umfassenden Qualtätsbegriff gesehen werden. Während das QoS Konzept hauptsächlich technische Kriterien verwendet und die Performance eines Systems evaluiert, ist die QoE vor allem der Wahrnehmung des Nutzers zugewandt. Im Zentrum stehen dabei die Begriffe „Quality“ und „Experience“. Experience wird dabei definiert als

> „ *(...) an individuals stream of perception and interpretation of one or multiple events.*“

und Quality als

> „ *(...) the outcome of an individuals comparison and judgement process. It includes perception, reflection about the perception and the description of the outcome. (...)* “

Somit kann QoE im Gegensatz zur QoS nicht nur durch physikalische Eigenschaften oder das Erfüllen gegebener Anforderungen erfasst werden. Wesentliche Bedeutung haben vielmehr die Wahrnehmung und die subjektive Beurteilung eines Individuums. Informationen über diesen rein kognitiven Vorgang lassen sich daher ausschließlich anhand von Beschreibungen des Nutzers gewinnen. Neben der Wahrnehmung des Nutzers spielt auch die Referenz, anhand der er sein vergleichendes Urteil fällt eine entscheidende Rolle. Abbildung 3.3 auf der nächsten Seite soll den Entscheidungsprozess verdeutlichen, der zum Qualitätsurteil eines Anwenders führt.

Aufbauend auf früheren Definitionen aus [6] und [38] wird QoE in [19] folgendermaßen definiert:

> „ *(...) the degree of delight or annoyance of the user of an application or service. It results from the fulfillment of his or her expectations with respect to the*

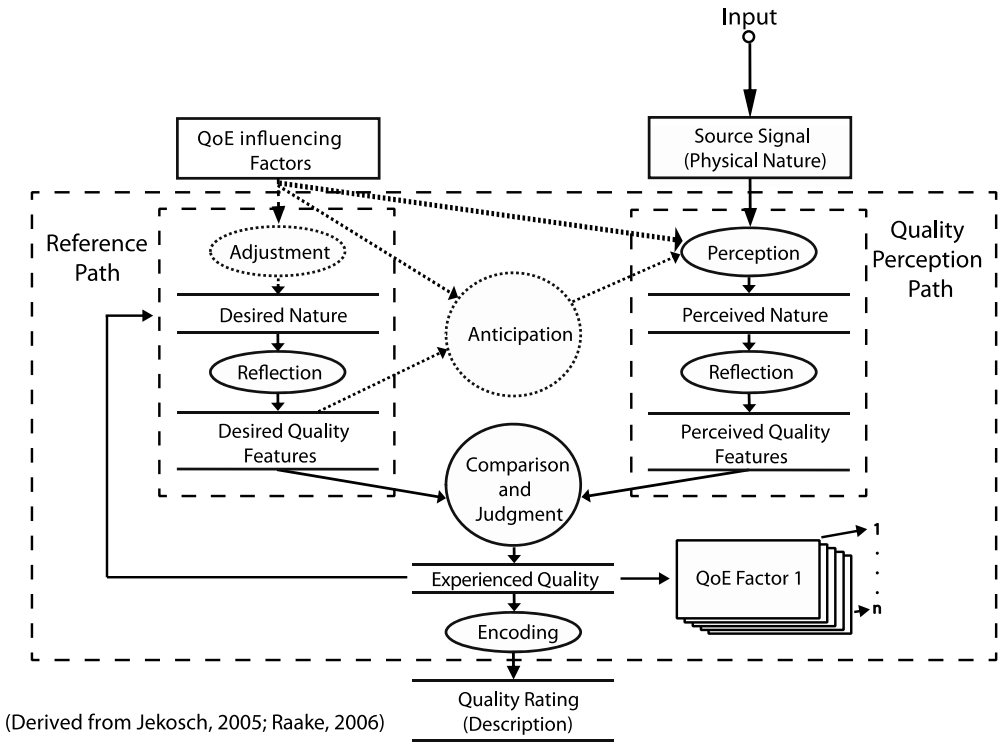


Abbildung 3.3.: Entstehungsprozess des Qualitätsurteils eines Anwenders, die Grafik ist [19] entnommen

> *utility and/or enjoyment of the application or service in the light of the users personality and current state.*“

Besonders hervorgehoben wird im weiteren Verlauf die Abgrenzung vom Begriff Performance, der nach [37] bestimmt ist durch „*The ability of a unit to provide the function it has been designed for*“. Erwähnt sei auch noch, dass nach [6] die Wahrnehmung im Umgang mit einem „*end to end system*“ maßgeblich für die QoE ist.

4. Subjektive Evaluierung

In diesem Kapitel der Arbeit soll nun ein umfassender Überblick über in der Literatur vorgestellte Methoden zur Evaluierung von Qualität durch die Befragung von Nutzern und die wesentlichen Schwierigkeiten dabei gegeben werden. Dieser Überblick soll gemeinsam mit der Analyse des Qualitätsbegriffes im vorigen Kapitel die theoretische Grundlage zur Entwicklung eines eigenen Evaluierungskonzepts für das in Kapitel 2 vorgestellte System bilden.

4.1. Planung des Experiments (Grundlagen)

Die Qualitäts-Evaluierung mit den Nutzern erfolgt im Rahmen eines Experiments. Daher soll zunächst auf die theoretischen Grundlagen zur Planung eines solchen Versuchs eingegangen werden. Soweit nicht anders gekennzeichnet bezieht sich Abschnitt 4.1 dabei auf das Buch „Perceptual Audio Evaluation - Theory Method & Application“ [10] von S. Bech und N. Zacharov.

4.1.1. Unabhängige und abhängige Variable

Die subjektive Wahrnehmung eines Versuchsteilnehmers wird meist von zahlreichen Faktoren beeinflusst. Diese Faktoren werden auch als Variable bezeichnet. Dabei wird zwischen folgenden Arten von Variablen unterschieden:

Versuchsvariable sind die Variablen, die im Rahmen des Experiments untersucht werden. Dabei wird zwischen unabhängigen und abhängigen Variablen unterscheiden. **Unabhängige Variable** sind diejenigen, die vom Versuchsleiter zur Gestaltung des Experiments variiert werden. **Abhängige Variable** nennt man hingegen diejenigen, die von den unabhängigen beeinflusst werden, also in erster Linie die Antworten der Versuchsteilnehmer.

Kontrollierte Variable sind dem Entwickler des Experiments bekannte Variable, die nicht zu den Versuchsvariablen zählen. Möglichkeiten solche Variablen zu kontrollieren sind, sie entweder konstant zu halten oder sie zu Randomisieren.

Störvariable sind unkontrollierte Variable, die meist mit den Versuchsvariablen kofundieren und somit das Versuchsergebnis verfälschen. Daher ist es von wesentlicher Bedeutung, den Einfluss der Störvariablen durch gutes Design des Experimens mög-

lichst gering zu halten. Dazu kann man entweder Versuchen, die Störvariablen zu kontrollierten Variablen zu machen oder sie zu Randomisieren.

Zufällige Variable sind schließlich alle Variablen, die in keine der bisher genannten Kategorien fallen. Sie tragen ebenfalls zum Fehler bei. Ihr Einfluss kann durch Randomisieren des Experiments gering gehalten werden.

Der subjektive Eindruck Y eines Versuchsteilnehmers (abhängige Variable) auf einen Stimulus (unabhängige Variable) besteht somit aus zwei Teilen:

$$Y = \mu + \varepsilon \tag{4.1}$$

Dabei wird μ von den unabhängigen und den kontrollierten Varablen und ε von den zufälligen und den Störvariablen bestimmt. Ziel bei der Planung eines Experiments sollte es also sein, den Term ε möglichst gering zu halten. Im Idealfall wird ε dann nur noch von Zufallsvariablen bestimmt. Die statistische Analyse des Experiments gibt dann eine Antwort auf die Frage, ob Veränderungen des subjektiven Eindrucks Y auf die beabsichtigte Variation der Stimuli zurückzuführen oder mit größerer Wahrscheinlichkeit zufälliger Natur sind.

4.1.2. Versuchsdesign

Beim Design des Versuchs gibt es, wie in [10] erläutert, zwei Ebenen. Zum einen das „Treatment Design“ und zum anderen die Zuordnung der Stimuli zu den Teilnehmern.
Beim Treatment Design gilt es, unabhängig von den Versuchspersonen, die Entscheidung zu treffen, welche Stimulus-Treatments im Experiment verwendet werden sollen und zu welchem Zweck. Ein Treatment wird dabei durch eine eindeutige Kombination unabhängiger Variablen definiert. Einen Sonderfall für das Treatment Design ist das sogenannte „Full Factorial Design“. Es enthält als Stimulus-Treatments alle möglichen Kombinationen von unabhängigen Variablen und ist aus statistischer Sicht die optimale Lösung. Für weitere Möglichkeiten sei auf Kapitel 6 in [10] verwiesen.
Bei der Zuordnung stellt sich in erster Linie die Frage, ob jedem Teilnehmer das vollständige Set an Treatments präsentiert werden kann. In diesem Fall spricht man von einem „Within-Subjects Design“. Diese Variante bietet den wesentlichen Vorteil, dass individuelle Wahrnehmungsunterschiede auf das ganze Set verteilt werden. Sollte diese Option aus zeitökonomischen oder logistischen Gründen nicht bestehen, so muss man zu einem „Between-Subjects Design“ greifen. Ansätze und Varianten dazu finden sich in [10].
Ein weiterer wichtiger Freiheitsgrad bei der Planung ist die Reihenfolge, in welcher den Teilnehmern die einzelnen Treatments präsentiert werden. Die Abfolge ist bei den meisten Experimenten von wesentlicher Bedeutung, da Veränderungen der subjektiven Wahrnehmung durch die bereits gehörten Treatments entstehen. Wenn also die Abfolge nicht explizit Gegenstand der Untersuchung ist, sollte man den Teilnehmern unterschiedliche Reihenfolgen präsentieren. Wählt man eine reine Zufallsreihenfolge, so macht man die

Abfolge zu einer Zufallsvariablen. Man kann aber auch Versuchen, diese Variable zu kontrollieren, indem man durch die Verwendung eines „Balanced Latin Square“ (BLS) sicherstellt, dass ein Treatment im Verlauf des Experiments jeweils nur ein einziges mal von einem bestimmten anderen gefolgt auftritt. Das BLS Design ist direkt nur auf eine gerade Anzahl an Treatments anwendbar. Bei ungeraden Zahlen erweitert man jede Zeile mit Ihrer umgekehrten Folge, verdoppelt damit allerdings auch den Aufwand.

4.2. Subjektive Bewertung und Kommunikation mit den Versuchsteilnehmern

Eine der größten Schwierigkeiten bei der subjektiven Qualitätsevaluierung stellt der Dialog mit den Nutzern dar. Es gilt dabei, zwischen dem Leiter des Experiments und den Probanden größtmögliches Einverständnis über den Inhalt des Versuchs herzustellen („was soll bewertet werden?“) sowie eine geeignete Form der Bewertungserfassung zu finden („wie soll bewertet werden?“). Letztere soll es dem Probanden ermöglichen, seine Wahrnehmung möglichst unverfälscht wiederzugeben.

4.2.1. Einweisung und Training

Wichtige Teile des Experiments, die helfen sollen Missverständnissen in der Kommunikation mit den Versuchspersonen vorzubeugen, sind die Einweisung und das Training.
Die Einweisung soll dem Teilnehmer seine Aufgabe im Hörversuch vermitteln. Dabei ist es sinnvoll - soweit man annehmen kann, die Wahrnehmung dadurch nicht zu beeinflussen - die Versuchsperson über das Ziel und den Aufbau des Experiments aufzuklären [10]. Die Einweisung besteht üblicherweise aus einem schriftlichen Dokument sowie einem Gespräch mit der Versuchsperson, das ausreichend Raum für Rückfragen geben sollte.
Das Training soll den Teilnehmern die Gelegenheit bieten, sich an das System zu gewöhnen und ihnen die wesentlichen Aspekte des Versuchs anhand von Beispielen verdeutlichen. Da Ablauf und Inhalt eines solchen Trainings in hohem Maße abhängig vom jeweiligen Experiment sind gibt es keine standardisierten Verfahren. In der Literatur wird aber an mehreren Stellen die Wichtigkeit einer sorgfältigen Planung dieser beiden Schritte betont und anhand von Beispielen illustriert [39], [10].

4.2.2. Bewertungsskalen

Ein weiterer wichtiger Bestandteil der Kommunikation bei Experimenten zur subjektiven Evaluierung sind Bewertungsskalen, auf denen die Versuchspersonen ihre Wahrnehmung bzw. ihr daraus resultierendes Urteil qualitativ oder quantitativ wiedergeben können. Die Antwortmöglichkeiten können dabei von binären Entscheidungen („Ja oder Nein“, „Besser oder Schlechter“) bis hin zu vergleichenden oder absoluten Bewertungen auf einer Skala mit kontinuierlichem Wertebereich reichen. Die Wahl einer geeigneten Skala ist ebenfalls

von wesentlicher Bedeutung um ein möglichst genaues Abbild der subjektiven Wahrnehmung der Probanden zu bekommen. Im folgenden sollen kurz einige gängige Skalen zur Qualitätsevaluierung vorgestellt werden.

5 Punkt Skala nach ITU-T P.800

Eine bei der Evaluierung von Telekonferenzsystemen häufig verwendete Skala ist die in ITU-T P.800 [1] beschriebene 5 Punkt Skala zur Erhebung des „Mean Listening-Quality Opinion Score“. Gegenstand von [1] ist die subjektive Bewertung der Übertragungsqualität. Systeme mit räumlicher Wiedergabe werden darin zwar nicht explizit erwähnt, eine Verwendung der Richtlinie zur Evaluierung eines solchen System ist aber durchaus denkbar. Ein wesentlicher Vorteil bei der Verwendung der 5 Punkt Skala ist, dass die Standardisierung den Vergleich mit anderen Studien erleichtert. Wie in Abbildung 4.1 zu sehen ist, handelt es sich bei der 5 Punkt Skala allerdings um eine diskrete Skala, worin ihre größte Schwäche liegt. Diskrete Skalen verursachen bei der Erfassung quantitativer Kriterien immer eine „Mapping Bias“ [44]. Das bedeutet, dass die tatsächliche wahrgenommene Qualität, die durchaus zwischen zwei der diskreten Werte liegen kann, einem der 5 Adjektive zugeordnet werden muss. So können unter Umständen Stimuli, die als geringfügig unterschiedlich wahrgenommen werden auf der Skala die gleiche Bewertung erhalten.

Abbildung 4.1.: Anwendungsbeispiel der ITU 5 Punkt Skala (Abbildung aus [44] entnommen)

ITU „Quality Skala“

Eine weitere häufig verwendete Skala ist die Quality Skala, wie sie in den Richtlinien ITU-R BS.1284-1 [3] und ITU-R BS.1534-1 [4] beschrieben wird. Sie ähnelt in ihren Grundzügen der 5 Punkt Skala und verwendet die gleichen Wortanker zur Beschreibung der Wahrnehmung. Der wesentliche Unterschied ist, dass sie im Gegensatz zu 5 Punkt Skala kontinuierlich ist. Abbildung 4.2 auf der nächsten Seite zeigt ein Anwendungsbeispiel der Skala aus [44].

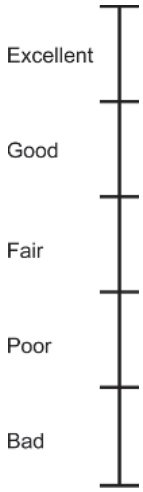


Abbildung 4.2.: DIe ITU Quality Skala aus [3], [4]) (Abbildung aus [44] entnommen)

Bodden Jekosch Skala

Die dritte und letzte Skala, die hier vorgestellt werden soll, ist die nach ihren Entwicklern benannte „Bodden Jekosch Skala“. Die Skala wurde in [16] zur Bewertung der Sprachübertragungsqualität von Telekonferenzsystemen eingeführt. Der Bedarf einer neuen Skala bestand darin, dass der ITU Standard - die in [1] empfohlene 5 Punkt Skala - die zuvor bereits in Abschnitt 4.2.2 erwähnten Schwächen hinsichtlich des Mapping Bias Effekt aufweist. Die Bodden Jekosch Skala ist daher wie die ITU Quality Skala kontinuierlich. Zusätzlich bietet sie die in Abbildung 4.3 sichtbaren künstlichen Extrembereiche, welche die Versuchspersonen dazu ermutigen sollen, auch die höchsten bzw. niedrigsten Attribute zu verwenden [37]. Die zusätzlichen Markierungen zwischen den 5 Attributen sollen der Skala eine Meterstab-Optik geben, die die Versuchspersonen dazu veranlassen soll, die Attribute als äquidistant einzustufen. Damit soll der in [44] erwähnte „Spacing Bias“ Effekt vermieden werden. Ein Nachweis über die Wirksamkeit dieser Änderungen steht allerdings laut [37] noch aus.
Beispiele zur Verwendung der Bodden Jekosch Skala finden sich in [37] und [40].

Abbildung 4.3.: Die Bodden Jekosch Skala (Abbildung aus [37]) entnommen)

4.2.3. Verwendung von Referenzsignalen

Im abschließenden Punkt zur Kommunikation mit den Versuchspersonen im Hörversuch soll im Folgenden noch auf die Verwendung sogenannter Referenzsignale eingegangen werden. Diese sollen dabei helfen, die Ergebnisse eines Experiments besser in einen weiteren Kontext setzen zu können. Wie in [44] beschrieben ist es nicht so einfach möglich, Qualität in einem absoluten Maß zu erfassen. Vielmehr basiert die subjektive Beurteilung, wie in Kapitel 3 beschrieben, zu großen Teilen auf Vergleichsprozessen. Referenzsignale bieten sowohl bei der subjektiven Bewertung als auch bei der Auswertung und Interpretation eines Experiments Orientierungspunkte.
Dabei sind zahlreiche Varianten denkbar. Zum Beispiel kann ein genormter Stimulus verwendet werden, der in vielen Studien zum selben Thema eingesetzt wird. Dadurch können die unterschiedlichen Ergebnisse stets in Relation zu diesem Referenzsignal betrachtet werden. Ein Beispiel dazu ist die Referenz (oder auch: Anker) des „MUSHRA“[1] Tests in [4]. Allgemein ist es laut [10] sinnvoll, Referenzen zu wählen, die entweder Extrempunkte der Skala oder mehrere über die Skala verteilte Ankerpunkte repräsentieren.

4.3. Testumgebung

Hinsichtlich des Versuchsraums verlangt die ITU für Hörversuche, bei denen nur Kopfhörer zur Wiedergabe verwendet werden, dass der Hintergrundgeräuschpegel mindestens dem ISO Noise Rating NR 15 entspricht [2]. Alle weiteren Spezifikationen gelten lediglich für Hörversuche, bei denen auch Lautsprecher verwendet werden. Für Hörversuche zur Telekommunikation wird von der ITU in [1] ein maximaler Hintergrundgeräuschpegel von $30 dB(A)$ sowie eine Raumgröße zwischen $30 - 120 m^3$ und eine Nachhallzeit unter $500 ms$ gefordert.
Andere Literatur, z.B. [10], widmet sich in erster Linie der Wiedergabe über Lautsprecher und gibt darüber hinaus noch allgemeine Hinweise, beispielsweise, dass für ausreichende Frischluftzufuhr zu sorgen ist.

[1] „Multi Stimulus with Hidden Reference Anchor“

5. Konzept und Experiment zur Evaluierung

In diesem Kapitel soll nun die Planung und Durchführung des Experiments zur Evaluierung des Systems beschrieben werden. Zunächst soll dazu auf das Konzept und das Design des Versuchs eingegangen werden, bevor anschließend der Ablauf sowie das gewählte Verfahren zur Auswertung der Daten vorgestellt werden.

5.1. Versuchsdesign

Im Abschnitt Versuchsdesign soll zunächst das allgemeine Evaluierungskonzept geschildert werden und anschließend detaillierter auf die im Experiment verwendeten Stimulus Treatments sowie deren Anordnung und die verwendete Bewertungsskala eingegangen werden.

5.1.1. Evaluierungskonzept

Ziel des Experiments war es, das Telekonferenzsystems hinsichtlich der in Abschnitt 3.5 vorgestellten Quality of Experience zu evaluieren, mit besonderem Augenmerk auf dem jeweiligen Beitrag der Komponenten Channel Assignement, binaurale Wiedergabe und Head-Tracking, die somit die unabhängigen Variablen des Experiments darstellen (vgl auch Abbildung 2.2).
Als Kriterium zur Evaluierung wurde daher der subjektive Gesamteindruck eines Probanden hinsichtlich seiner Zufriedenheit mit der gehörten Präsentation einer simulierten Telekonferenzsituation gewählt.
Neben der QoE in einer gewöhnlichen Konferenzsituation sollte zudem noch in einem zweiten Teil des Hörversuchs die QoE in einer besonders unübersichtlichen Konferenzsituation evaluiert werden, wobei in erster Linie die Frage geklärt werden sollte, wie der Channel Assignement Algorithmus in solch einer Situation bewertet wird und ob trotz der auftretenden Artefakte noch die räumlich getrennte Wiedergabe der Sprecher bevorzugt wird. Da vor dem Experiment nicht abgeschätzt werden konnte, welche Veränderungen die als stark einzustufenden Artefakte dieses zweiten Teils in der Wahrnehmung der Versuchspersonen hervorrufen würden, wurde dieser Teil in allen Durchläufen in einem separaten Block am Ende des Experiments platziert.
Als Refernzen (vgl. Abschnitt 4.2.3) wurden für das Channel Assignement die ideale Trennung des Headset Szenarios aus Abschnitt 2.1 auf Seite 11 und für das Head-Tracking das AR Tracking System eingeplant. Beide können als bestmögliche Optionen im jeweili-

Ch-Ass.	Algorithmus 4												Headset											
Tracking	AR				USB				off				AR				USB				off			
HRTF	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4

Tabelle 5.1.: Überblick über die 24 Stimulus Treatments im ersten Teil des Hörversuchs (Zuordnung der HRTFs: 1=KEMAR, 2=DOMISO, 3=PLSR, 4=Individuell)

Tracking	AR Tracking			
HRTF	KEMAR			
Ch-Ass.	Algorithmus 4		Headset	
Sprecherverteilung	4 Positionen	1 Position	4 Positionen	1 Position

Tabelle 5.2.: Überblick über die 4 Stimulus Treatments für den zweiten Teil des Hörversuchs

gen Bereich eingestuft werden. Für die binaurale Wiedergabe wurde keine Referenz eingesetzt, da es ja Ziel des Experiments war, herauszufinden, welche HRTF Datenbank am besten bewertet wird. Von einer mono Wiedergabe als unteres Limit wurde Abstand genommen, um das Risiko zu vermeiden, dass der starke Unterschied zwischen binauraler und mono Wiedergabe die Probanden für die Unterschiede zwischen den verschiedenen HRTF Datensätzen desensibilisiert [8].

5.1.2. Stimulus Treatments

Mit den möglichen Optionen für das Channel Assignement (2 Varianten), das Head-Tracking (3 Varianten, inkl. Off-Zustand) und die HRTF Datenbank (4 Varianten) ergibt sich eine Gesamtanzahl von $2 \cdot 3 \cdot 4 = 24$ möglichen Stimulus Treatments, die in Tabelle 5.1 aufgelistet sind. Da diese Menge für den Hörversuch als bewältigbar eingestuft wurde, fiel die Entscheidung auf das in Abschnitt 4.1.2 auf Seite 28 vorgestellte Full Factorial Design. Da auch davon ausgegangen werden konnte, das mit dem in Abschnitt 5.1.3 auf der nächsten Seite beschriebenen Stimulus die Gesamtdauer des Experiments für einen einzelnen Probanden auch bei Verwendung des vollständigen Stimulus Sets nicht zu lang werden würde, wurde des Weiteren ein Within Subjects Design gewählt (vgl. ebenfalls Abschnitt 4.1.2).
Für den zweiten Teil des Versuchs sollte nur zwischen den zwei Channel Assignement Varianten und der räumlich getrennten im Gegensatz zu einer Wiedergabe, bei der alle Sprecher auf der selben Position platziert sind, unterschieden werden. Daraus ergaben sich die vier in Tabelle 5.2 gelisteten Treatments für den zweiten Teil des Hörversuchs.
Die Treatments für den ersten und den zweiten Teil des Versuchs wurden jeweils in separaten Balanced Latin Squares, wie im folgenden Abschnitt 5.1.4 beschrieben, angeordnet.

Lautsprecher	KS C5 Tiny
Audio interface	Hammerfall DSP Multiface II
Vorverstärker	Focusrite Saffire Pro 40
Rechner	Lenovo Thinkpad (Ubuntu)

Tabelle 5.3.: Verwendetes Equipment zur Aufnahme der Stimuli im Videolab des LDV

5.1.3. Stimuli

Da davon auszugehen ist, dass ein Telekonferenzsystem in der überwiegenden Mehrheit der Fälle zur Übertragung von Sprache genutzt wird, ist es nur konsequent, bei der Evaluierung Sprachsignale zu verwenden. Dies ist auch in der Literatur so üblich [40].
Die für den Hörversuch verwendeten Daten stammen aus einem Meeting Korpus, der speziell zur Bewertung von Telekonferenzsystemen für mehrere Teilnehmer erstellt wurde. Die Daten wurden dankenswerter Weise von Herrn Janto Skowronek von den T-Labs zur Verfügung gestellt. Eine ausführliche Dokumentation zu den Aufnahmen findet sich in [47].
Um den Versuchspersonen ausreichend Zeit zu geben, sich in die Konferenzsituation hineinzuversetzen wurde ein Ausschnitt mit einer Länge von 58 Sekunden und vier Konferenzteilnehmern gewählt. Der Ausschnitt stammt aus Szenario 5 im Korpus aus [47]. In diesem Szenario gibt es zwar eigentlich sechs Teilnehmer, in dem gewählten Ausschnitt kommen jedoch nur vier davon zu Wort. Der Ausschnitt beginnt bei 6:24 min der editierten Version des Szenario 5. Die Daten, die für die Einführung zum Experiment sowie das Training der Modelle für die Sprechererkennung [48] verwendet wurden sind am Beginn des Szenario 5 zu finden. Dort stellt sich jeder der Sprecher ca 10 Sekunden lang vor.
Zur Bearbeitung mit dem in Abschnitt 2.1 vorgestellten Algorithmus zum Channel Assignement mussten die Daten zunächst mit dem Mikrophon Array des Systems aufgenommen werden. Die Aufnahmen fanden im Videolabor am LDV statt. Abbildung 5.1 auf der nächsten Seite zeigt den Aufbau. Die vier Lautsprecher befanden sich dabei im Abstand von $1.30m$ mit einem Elevationswinkel von 20° oberhalb der Konferenzspinne. Der Winkelversatz zwischen den Lautsprechern war jeweils 90°. Eine Liste des verwendeten Equipments findet sich in Tabelle 5.3.

5.1.4. Präsentationsreihenfolge (BLS)

Die Präsentationsreihenfolge der Treatments kann einen wesentlichen Einfluss auf die Bewertung haben (vgl. Abschnitt 4.1.2 auf Seite 28). Wie bereits in den Abschnitten 3.2 auf Seite 20 und 3.5 auf Seite 24 erwähnt, spielen Vergleiche bei der Bewertung eine wesentliche Rolle. Daher ist davon auszugehen, dass vor allem das zuletzt gehörte Beispiel einen starken Einfluss auf die Bewertung des nächsten hat. Außerdem kann man annehmen, dass sich die Wahrnehmung der Versuchspersonen im Laufe des Experiments ändert und sie einen sichereren Umgang mit ihren subjektiven Bewertungskriterien erlangen.
Daher wurde für das Experiment in dieser Arbeit ein Balanced Latin Square Design ge-

Abbildung 5.1.: Aufbau zur Aufnahme der Stimuli im Videolabor des LDV

$$\begin{array}{cccccc} 1 & n & 2 & n-1 & \dots & n/2+1 \\ 2 & 1 & 3 & n & \dots & n/2+2 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ n & n-1 & 1 & n-2 & \dots & n/2 \end{array}$$

Abbildung 5.2.: Aufbau des Balanced Latin Square

wählt. Dieses Design stellt sicher, dass jeder Proband mit einem anderen Beispiel beginnt und über das ganze Experiment nie zwei Teilnehmer zwei gleiche Treatments in direkter Abfolge nacheinander zu hören bekommen. Das BLS ist dabei so aufgebaut, dass die erste Spalte alle n Treatments in aufsteigender Reihenfolge enthält. Die zweite Spalte beginnt mit Treatment n, die dritte mit Treatment 2, die vierte mit $n-1$ usw. Dem liegt die einfache Überlegung zugrunde, dass jede Spalte aus der gleichen Folge bestehen soll, die immer um einen anderen Abstand zu der ihr links benachbarten Spalte verschoben wird. Abbildung 5.2 soll den Aufbau des BLS verdeutlichen. Wie aus der Abbildung noch einmal deutlich wird, lässt sich solch ein BLS nur für gerade Werte n bilden.

Betrachtet man noch einmal die in Abschnitt 5.1.2 auf Seite 34 beschriebenen Treatments für den ersten und zweiten Teil, so ergibt sich für das Experiment ein BLS der Größe $n = 24$ für den ersten Teil und 6 kleine BLS der Größe $n = 4$ überinander für den zweiten Teil. Eine Zeile dieser Matrix ist nun die Treatment Reihenfolge für einen Probanden.

5.1.5. Verwendete Skala

Zur Bewertung wurde die in Abschnitt 4.2.2 auf Seite 31 vorgestellte Bodden Jekosch Skala gewählt. Da das Ziel der vorliegenden Arbeit in erster Linie ein Vergleich verschiedener Optionen bei der Entwicklung des Systems ist, spielt die Vergleichbarkeit mit anderen Studien lediglich eine untergeordnete Rolle. Zudem verwendet die Bodden Jekosch Skala die selben Wortanker wie die Skalen der ITU. Die Vorteile der erweiterten Extrembereiche sowie der Linearisierung durch die zusätzlichen Markierungen wurden daher als wichtiger erachtet, auch wenn diese nicht abschließend nachgewiesen sind. Ausschlaggebend für die Entscheidung war nicht zuletzt auch die Korrespondenz mit Herrn Janto Skowronek von den Telekom Innovation Laboratories (T-Labs).

5.2. Versuchsaufbau

Das Experiment fand im Audiolabor des LDV statt. Die Probanden wurden in der Mitte des Raumes an einen Tisch gesetzt, auf dem sich der Rechner befand. Auf diesem wurde sowohl die Datenverarbeitung für das 3-D Audio Display sowie die Datenerfassung zur Evaluierung über ein eigens für das Experiment implementiertes Interface durchgeführt. Der Rechner war leicht erhöht positioniert, sodass die USB Kamera, die direkt darüber

Abbildung 5.3.: Ein Proband vor dem Aufbau zur Durchführung des Experiments im Audiolabor des LDV

Rechner	Lenovo Thinkpad (Ubuntu)
Audio Interface	Roland UA 25 EX
Kopfhörer	Beyerdynamic DT 990 Pro (offen)

Tabelle 5.4.: Verwendetes Equipment beim Hörversuch

angebracht war, sich auf Augenhöhe der Probanden befand. Um die unterschiedliche Körpergröße der Probanden auszugleichen wurde zu Beginn des Experiments die Höhe des Stuhls entsprechend eingestellt. Messinstrumente kamen dabei nicht zum Einsatz, da das USB Kamera Tracking unter realistischen Bedingungen getestet werden sollte. Die Audiowiedergabe erfolgte auf einem Stereo Kopfhörer, auf dessen Bügel der Tracking Body für das AR-Tracking System angebracht war. Die drei Infararotkameras befanden sich in den zwei oberen Raumecken vor sowie links hinter dem Probanden in einem Abstand von ca. 2.5m vom Tracking Body. In Abbildung 5.3 sieht man einen Probanden am Versuchsaufbau. Tabelle 5.4 listet das verwendete Equipment auf.
Bezüglich der in Abschnitt 4.3 erwähnten Anforderungen an die Testumgebung erfüllt das Audiolabor des LDV die Vorgaben aus ITU-T P.800 [1] hinsichtlich Volumen und Nachhallzeit. Der Geräuschpegel bei laufendem Betrieb der Infrarot Kameras wurde mangels eines entsprechenden Messgerätes nicht überprüft, kann aber bei Bedarf im Nachhinein noch ergänzt werden.

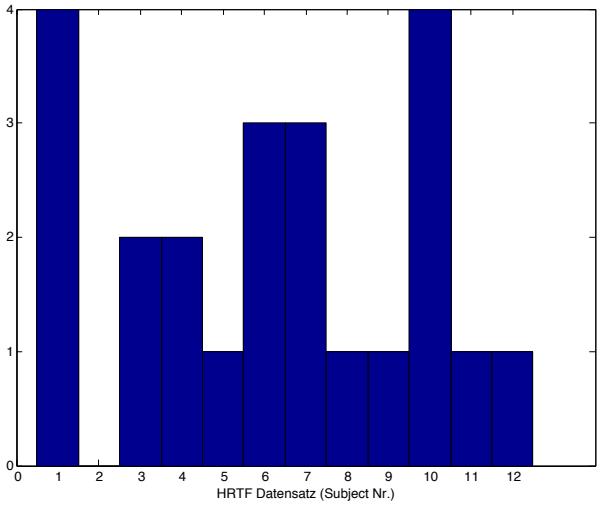


Abbildung 5.4.: Histogramm zum Ausgang des DOMISO HRTF Selection Verfahrens

5.3. Ablauf des Experiments

In diesem Abschnitt soll nun der Ablauf des Experiments mit den Probanden beschrieben werden. Jeder Proband musste dabei die HRTF Selection, eine Einweisung und ein kurzes Training sowie den eigentlichen Hörversuch absolvieren. Die Aufnahme der individuellen HRTFs war bereits abgeschlossen.

5.3.1. HRTF Selection (DOMISO)

Der erste Teil des Experiments war das HRTF Selection Verfahren, wie es in Kapitel 2.2.3 auf Seite 16 beschrieben ist. Die schriftliche Einweisung, die die Probanden zu Beginn des Versuchs vorgelegt bekamen findet sich in Anhang A auf Seite 61. Zusätzlich zu der schriftlichen Einweisung wurden die Probanden noch darauf hingewiesen, dass sie für die Bewertung in paarweisen Vergleichen neben den in der Einführung vorgeschlagenen auch jegliche weiteren, ihnen sinnvoll erscheinenden, Kriterien heranziehen können. Ausserdem wurde noch erwähnt dass die Kreisbewegung des Rosa Rauschens auf der Elevationsebene von 0° verlaufen sollte (das stand nicht explizit auf der Einweisung).
Abbildung 5.4 zeigt das Ergebnis des Verfahrens. Die Höhe der Balken des Histogramms repräsentiert, wie oft jeder der 12 zur Wahl stehenden HRTF Datensätze gewählt wurde.

5.3.2. Einweisung und Training

Die schriftliche Einweisung zum Hörversuch bestand aus zwei Einweisungsblättern, die sich in Anhang B auf Seite 64 und Anhang C auf Seite 65 finden. Neben den schriftlichen

Hinweisen wurden die Probanden angewiesen, ihr Blickfeld auf den Monitor des Rechners oder zumindest auf die ihnen gegenüberliegende Wand gerichtet zu halten, um zu gewährleisten, dass sich ihr Gesicht zu jedem Zeitpunkt im Blickfeld der USB Kamera befindet. Über die Funktion der USB Kamera wurden sie aber im unklaren gelassen. Insbesondere wurden die Probanden noch einmal darauf hingewiesen, dass die Adjektive auf der Skala ihre subjektive Zufriedenheit bei der Benutzung eines Telekonferenzsystems wiedergeben sollen. Daher sollten sie versuchen, sich so weit wie möglich als Zuhörer in die Telekonferenzsituation hineinzuversetzen. Ebenfalls hingewiesen wurden die Probanden auf die Tatsache, dass der als Sprecher 2 gekennzeichnete Konferenzteilnehmer im ersten Teil des Hörversuchs nur wenige Worte sagt, jedoch in den letzten vier Hörbeispielen als Unterbrecher aktiv wird, da dies bei den ersten Probanden teilweise zu Verwirrung führte.
Das Training sollte vier Aufgaben erfüllen. Erstens sollte es den Probanden die binaurale Wiedergabe des Systems demonstrieren. Dazu wurde ihnen zu Beginn der Trainingsausschnitt (die in Abschnitt 5.1.3 erwähnte „Vorstellungsrunde“) einmal mit einer herkömmlichen Mono Wiedergabe und im Anschluss einmal zum Vergleich binaural aufbereitet vorgespielt.
Zum zweiten sollte das Training die Versuchspersonen für die binaurale Wiedergabe mit Head-Tracking sensibilisieren. Dazu wurde als drittes Beispiel der Trainingsausschnitt mit aktivem AR Tracking präsentiert und die Versuchspersonen aufgefordert ihren Kopf zu bewegen und auf die Quellpositionen zu achten. Die Hörbeispiele 4 und 5 waren eine Wiederholung der Beispiele 2 und 3, um den Probanden etwas Zeit zu geben, sich an die binaurale Wiedergabe und das Head-Tracking zu gewöhnen.
Das dritte Anliegen im Training war es, die Teilnehmer für die Artefakte, die durch den Channel Assignement Algorithmus entstehen können zu sensibilisieren. Dazu wurde eine Version des Trainingsausschnitts aufgenommen, bei der der zweite und der dritte Sprecher sich um ca. $0.2s$ und der dritte und der vierte Sprecher sich um ca. $3s$ überlappen, was bei einer Verarbeitung durch den Algorithmus zu leichten bzw. stärkeren Artefakten führt.
Der vierte und letzte Punkt des Trainings, war die Einführung des Interface zur Evaluierung.. Dazu wurde den Probanden einmal das Dialogfenster mit der Bodden Jekosch Skala gezeigt und die Bedienung des Sliders mit der Maus demonstriert.
Zur binauralen Wiedergabe wurden während dem gesamten Training die KEMAR HRTFs verwendet, um allen Teilnehmern die gleichen Voraussetzungen zu geben. Zum Tracking wurde das als ideal eingestufte AR Tracking verwendet. Tabelle 5.5 auf der nächsten Seite zeigt noch einmal den Ablauf des Trainings.

5.3.3. Hörversuch

Im Verlauf des Hörversuchs bekamen die Probanden jeweils ein Stimulus-Treatment vorgespielt, welches sie direkt im Anschluss bewerten sollten. Diese Prozedur wurde solange wiederholt, bis alle Treatments in der durch die Balanced Latin Squares gegebenen Reihenfolge abgearbeitet waren. Die Probanden hatten auch die Option, das aktuelle

	Stimulus	Binaurale WIedergabe	Head-Tracking
Beisoiel 1	Vorstellungsrunde	off	off
Beispiel 2	Vorstellungsrunde	on	off
Beispiel 3	Vorstellungsrunde	on	on
Beispiel 4	Vorstellungsrunde	on	off
Beispiel 5	Vorstellungsrunde	on	on
Beispiel 6	Vorstellungsrunde mit Artefakten	on	on
Beispiel 7	Demonstration des Evaluierungs Interface		

Tabelle 5.5.: Schematischer Ablauf des Trainings

Hörbeispiel zu wiederholen, von dieser wurde jedoch kein Gebrauch gemacht.

5.4. Präsentation und Auswertung der Daten

In diesem Abschnitt sollen abschließend noch die gewählten Verfahren zur Präsentation und zur Auswertung der Daten aus dem Experiment vorgestellt werden.

5.4.1. Präsentation (Boxplots)

Um eine erste Übersicht über die Daten zu liefern sollen in dieser Arbeit die z.B. in [24] vorgestellten Boxplots verwendet werden. In Abbildung 5.5 ist ein Beispiel zu sehen.
Die rote Linie innerhalb der Box markiert den Median der Stichprobe. Die Box umfasst den Interquartilbereich, in dem sich 25% der Daten oberhalb so wie 25% der Daten unterhalb des Medians (also insgesamt 50% der Daten) befinden. Die Whisker außerhalb der Box haben die 1.5-fache Länge des jeweiligen 25%-Quartils. Daten, die außerhalb des Whisker Bereichs liegen werden als Ausreißer bezeichnet. Der taillierte Bereich der Box markiert das 95%-Konfidenz Intervall. Zusätzlich soll in dieser Arbeit noch der arithmetische Mittelwert durch das Raute-Symbol gekennzeichnet werden.

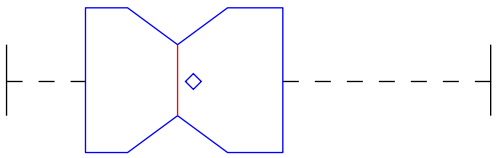


Abbildung 5.5.: Boxplot

5.4.2. Auswertung (ANOVA)

Zur Auswertung des Experiments wurde die in [10] vorgeschlagene „Analysis of Variance“ (ANOVA) gewählt.
Die ANOVA ist ein statistisches Werkzeug zum Vergleich mehrerer experimenteller Beobachtungen verschiedener Treatments - oder genauer: der Mittelwerte dieser Stichproben. Dabei soll überprüft werden, ob Unterschiede zwischen den Stichproben zufälliger Natur sind, oder auf einen systematischen Effekt zurück schließen lassen. Dazu wird kurz gesagt die Varianz innerhalb der Stichproben mit der Varianz zwischen den Stichproben verglichen. Die ANOVA produziert dazu die sogenannte F-Statistik, welche die systematische Varianz mit der zufälligen (Fehler-)Varianz in Relation bringt [24].
Ausgangspunkt bei der ANOVA ist stets die Nullhypothese, die besagt, dass es keinen signifikanten Unterschied zwischen den Mittelwerten der betrachteten Stichproben gibt. Steigt die systematische Varianz schneller als die zufällige und somit auch die F-Statistik, so steigt die Chance, dass man die Nullhypothese verwerfen kann. Der p-Wert der ANOVA gibt Auskunft über die Wahrscheinlichkeit, dass die vorliegende Beobachtung bei wahrer Nullhypothese auftritt. Gängige Schwellwerte für p sind $p = 0.05$ oder $p = 0.01$ [24]. Unterschreitet p diesen Schwellwert, so bezeichnet man den Unterschied zwischen den Stichproben als signifikant. Im Rahmen dieser Arbeit wird der Schwellwert $p = 0.05$ verwendet.
Vergleicht man mehr als 2 Mittelwerte und kann die Nullhypothese verwerfen, so liefert einem die ANOVA allerdings nur das Ergebnis, dass es einen signifikanten systematischen Unterschied zwischen den Mittelwerten der Stichproben gibt, aber nicht zwischen welchen genau. Diese Information erhält man anhand vorab geplanter paarweiser Vergleiche [24]. oder sogenannter „Post Hoc“-Verfahren, wie dem „Least SIgnificant Difference Test“, der Bonferroni Korrektur oder dem Tukey Verfahren [24].
Für statistische Auswertung der Daten wurden die jeweiligen Implementierungen der Verfahren aus der MATLAB Statistics Toolbox verwendet.

6. Ergebnisse des Experiments

An dem Experiment nahmen insgesamt 20 Personen Teil. Es handelt sich dabei um Studierende und Doktoranden der TU München (4 weibliche, 16 männliche, zwischen 20 und 30 Jahren).
Eine erste Übersicht über die in dem Hörversuch ermittelten Daten zu allen 28 verwendeten Treatments findet sich in Abbildung D.1 auf Seite 68. Da die Daten so noch recht unübersichtlich sind, soll im weiteren Verlauf dieses Kapitels eine eingehendere Analyse hinsichtlich der unabhängigen Variablen des Experiments erfolgen.

6.1. Vergleich der Channel Assignement Varianten

Zunächst sollen hier die zwei Varianten des Channel Assignement miteinander verglichen werden. Dazu wurden alle Beobachtungen, die im ersten Teil des Versuchs unter Verwendung von Algorithmus 4 gemacht wurden zu einer Stichprobe zusammengefasst und alle Beobachtungen für das Headset Szenario zu einer zweiten. Abbildung 6.1 auf der nächsten Seite zeigt den Boxplot der zwei Stichproben und zusätzlich die als blaue Diamanten eingezeichneten arithmetischen Mittel. Die ANOVA liefert beim Vergleich der Proben einen p-Wert von $p = 6.05 \cdot 10^{-6}$. Die Nullhypothese kann somit verworfen werden. Der Vergleich der Mittelwerte ergibt also, wie erwartet, einen Vorteil für die ideale Trennung. Dieser ist jedoch nicht allzu groß. Während das Headset Szenario im Mittel knapp mit „gut“ bewertet wird, landet der Algorithmus 4 bei einem soliden „ordentlich“, was unter Berücksichtigung der erschwerten Bedingungen bei der Aufnahme mehrerer Sprecher durch die Konferenzspinne ein gutes Ergebnis ist.

6.2. Vergleich der Optionen zum Head-Tracking

Nun soll ein Vergleich der Optionen zum Head-Tracking vorgenommen werden. Dabei wurden alle Beobachtungen, die unter Verwendung jeweils einer der drei Optionen (AR Tracking, Webcam, „Off“) im Experiment gemacht wurden, zu einer Stichprobe zusammengefasst. Der Boxplot in Abbildung 6.2 auf Seite 45 zeigt die Verteilung der Daten. Der p-Wert der ANOVA von $p = 1.04 \cdot 10^{-6}$ zeigt, dass es mit hoher Wahrscheinlichkeit einen signifikanten Unterschied zwischen den Mittelwerten gibt. Um herauszufinden, ob es signifikante Unterschiede im paarweisen Vergleich der Mittelwerte gibt, soll das Least Siginificant Difference (LSD) Verfahren herangezogen werden. Das Ergebnis ist in Abbildung 6.3 auf Seite 46 zu sehen. Da keines der drei Intervalle die 0 beinhaltet, kann man

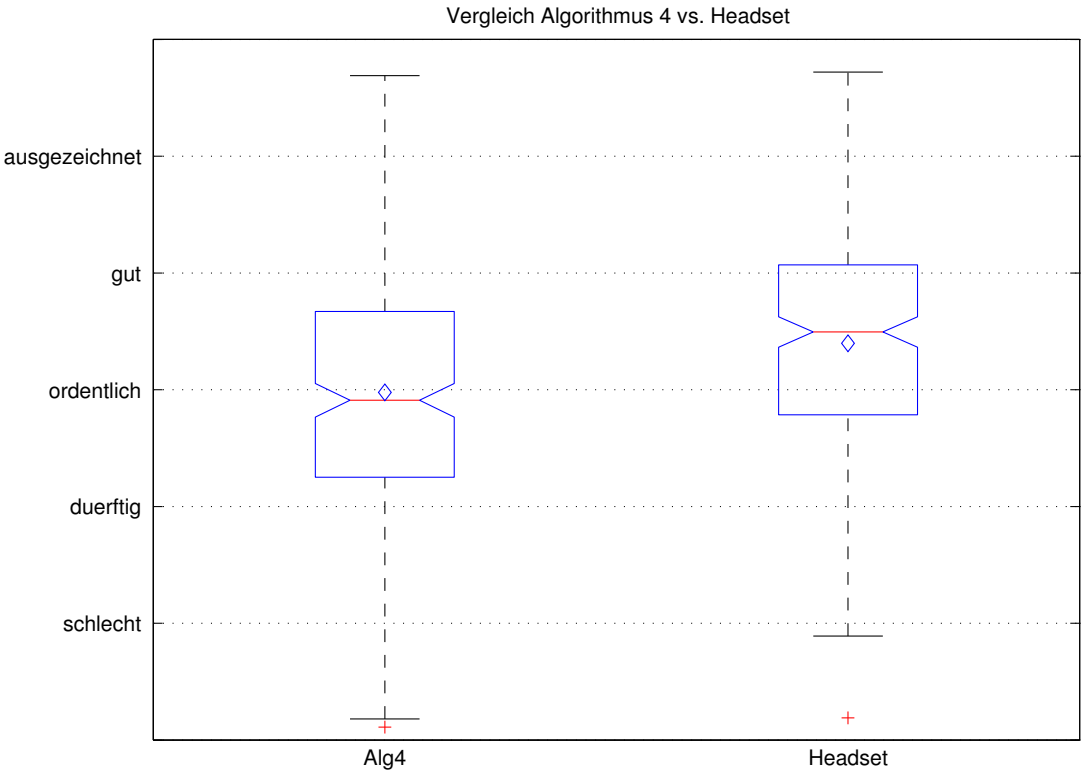


Abbildung 6.1.: Vergleich des Channel Assignement mit Algorithmus 4 und dem ideal getrennten Headset Szenario

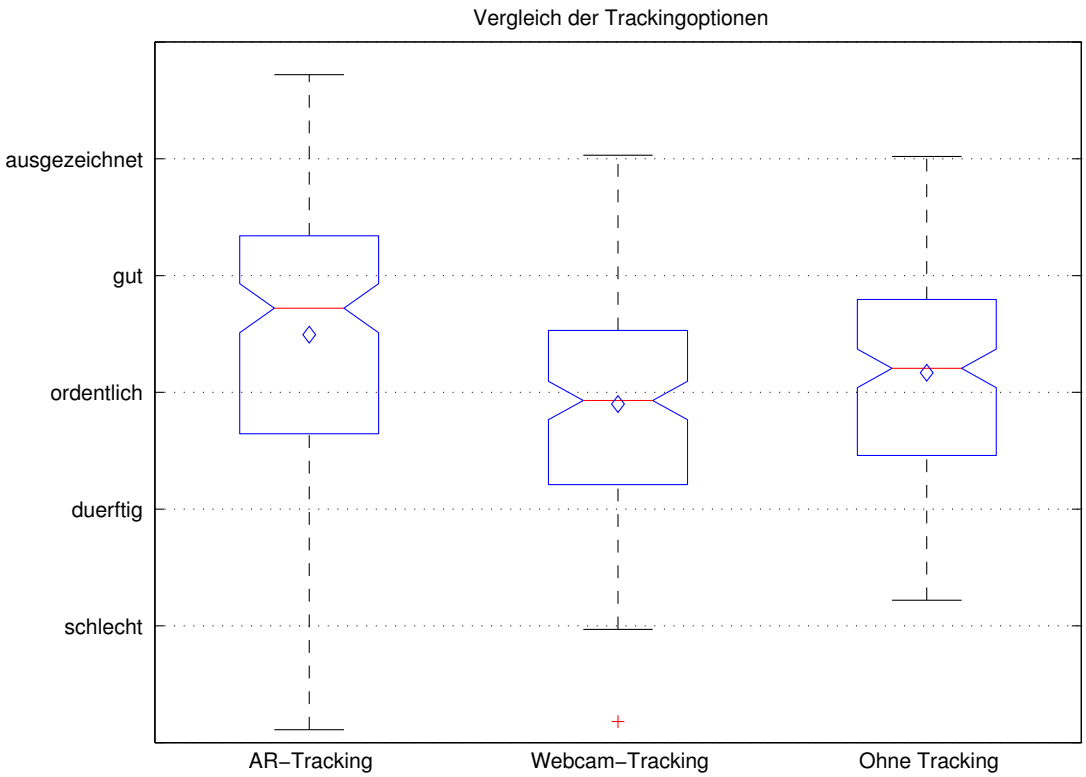


Abbildung 6.2.: Vergleich der drei im Experiment verwendeten Head-Tracking Optionen

1	2	37.44	59.38	81.32
1	3	10.66	32.60	54.54
2	3	−48.72	−26.78	−4.84

Abbildung 6.3.: Ergebnis des Least Siginficant Difference Verfahrens für den Vergleich der Head-Tracking Optionen

also alle Unterschiede zwischen den drei Mittelwerten als signifikant ansehen.
Es gibt demnach einen erwartungsgemäßen Vorteil für das Head-Tracking mit dem AR Tracking System. Etwas überraschend wird jedoch das Tracking mit der USB Kamera sogar schlechter als der Fall ohne Head Tracking bewertet.

6.3. Vergleich der HRTF Datenbanken

In diesem Abschnitt sollen nun die vier Verfahren zur Anpassung der HRTF Datenbank an den Nutzer miteinander verglichen werden. Abbildung 6.4 auf der nächsten Seite gibt eine erste Übersicht über die Daten. Auf den ersten Blick scheinen die drei Verfahren, die direkt „gemessene“ HRTFs verwenden, etwa gleichauf zu liegen. Die PLSR scheint etwas schlechter bewertet zu sein. Aus der ANOVA ergibt sich ein p-Wert von $p = 9 \cdot 10^{-4}$. Es ist also zumindest ein Mittelwert signifikant unterschiedlich von den anderen. Aufschluss über die paarweisen Vergleiche gibt das Ergebnis des LSD Tests in Abbildung 6.5 auf der nächsten Seite. Es gibt demzufolge über den gesamten ersten Teil des Experiments gesehen nur zwei paarweise Vergleiche der HRTF Datenbanken, die einen signifikant unterschiedlichen Mittelwert bei der Evaluierung produzieren, nämlich die Kombinationen KEMAR HRTF vs. PLSR und Individuelle HRTF vs. PLSR. Betrachtet man allerdings die Kombination DOMISO vs. PLSR (2 und 3 in Abbildung 6.5 auf der nächsten Seite), so fällt auf, dass es sich hier um einen Grenzfall handelt. Tendenziell lässt sich also die Aussage treffen, dass die durch die PLSR errechneten HRTFs von den Nutzern schlechter bewertet wurden, als die direkt aus einer Messung gewonnenen.

Da die Ergebnisse der Evaluierung bezüglich der HRTF Datenbanken soweit noch nicht allzu ergiebig sind, soll im Folgenden ein genauerer Blick auf die Ergebnisse bei Berücksichtigung der verschiedenen Head-Tracking Verfahren geworfen werden. Die Aufteilung der Daten nach den verschiedenen Trackingverfahren erfolgt insbesondere aufgrund der bekannten positiven Auswirkungen von Head-Tracking auf die binaurale Wiedergabe [11].

6.3.1. Vergleich der HRTF Datenbanken bei Verwendung des AR Tracking Systems

Zunächst sollen die vier Datenbanken bei Verwendung des als ideal eingestuften AR Tracking Systems betrachtet werden. Der Boxplot findet sich in Abbildung 6.6 auf Seite 48. Aus der ANOVA ergibt sich ein p-Wert von $6 \cdot 10^{-3}$. Die Ergebnisse des LSD Verfahrens

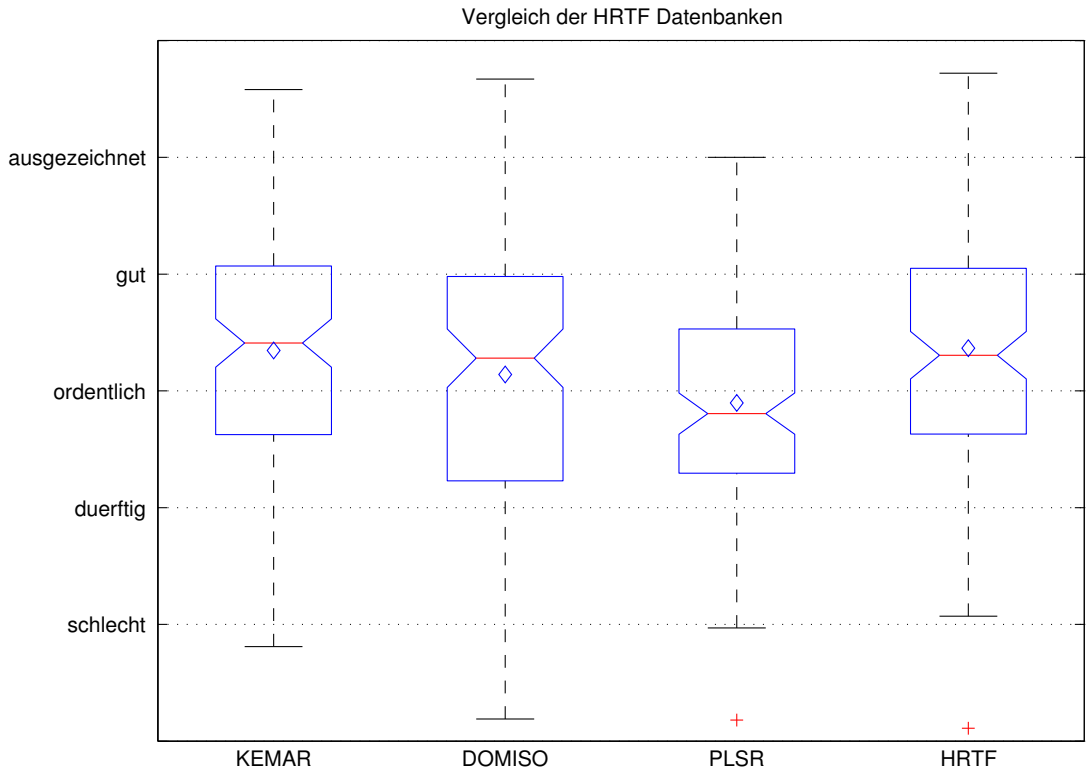


Abbildung 6.4.: Vergleich der vier HRTF Datenbanken

1	2	−5.14	20.52	46.17
1	3	19.24	44.90	70.56
1	4	−27.66	−2	23.66
2	3	−1.27	24.38	50.04
2	4	−48.17	−22.52	3.14
3	4	−72.56	−46.90	−21.24

Abbildung 6.5.: Ergebnis des Least Siginficant Difference Verfahrens für den Vergleich der vier HRTF Datenbanken

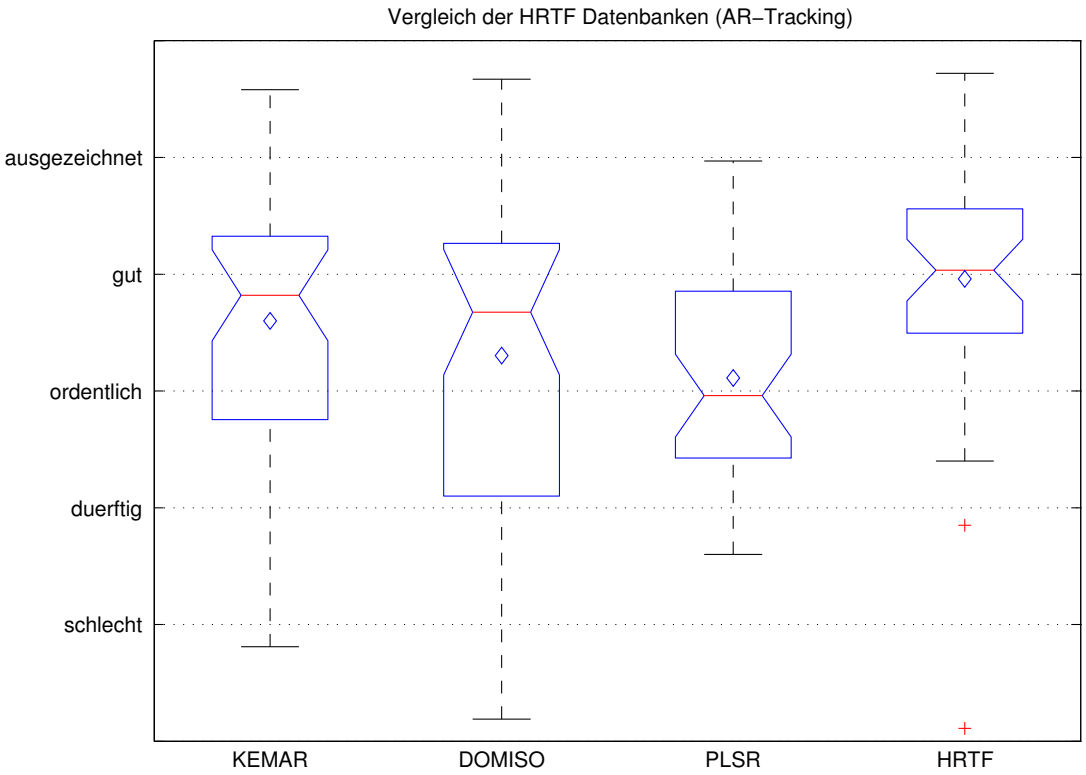


Abbildung 6.6.: Vergleich der vier HRTF Datenbanken bei Verwendung des AR Tracking Systems

finden sich in Abbildung 6.7 auf der nächsten Seite. Das Ergebnis wird hier insofern klarer, als dass die individuellen HRTFs signifikant besser abschneiden, als die HRTFs aus dem DOMISO Verfahren und die via PLSR errechneten. Ein Vorsprung zu den KEMAR HRTFs ist zwar vorhanden jedoch nicht signifikant.

6.3.2. Vergleich der HRTF Datenbanken bei Verwendung des Webcam Tracking Systems

Als nächstes sollen die HRTF Datenbanken unter Verwendung des Webcam Trackings betrachtet werden. Die Boxplot Übersicht findet sich in Abbildung 6.8 auf Seite 50. Die ANOVA liefert hier ein p-Wert von $p = 0.39$, der deutlich über dem geforderten Schwellwert liegt. Daher ist davon auszugehen, dass die Unterschiede zwischen den Mittelwerten nicht

1	2	−20.08	29.73	79.53
1	3	−0.93	48.87	98.68
1	4	−85.73	−35.92	13.88
2	3	−30.66	19.15	68.96
2	4	−115.46	−65.65	−15.84
3	4	−134.61	−84.80	−34.99

Abbildung 6.7.: Ergebnis des Least Siginficant Difference Verfahrens für den Vergleich der vier HRTF Datenbanken bei Verwendung des AR Tracking Systems

systematischer Natur sind.
Dieses Ergebnis lässt sich so interpretieren, dass die allgemein schlechter bewerteten Effekte des Webcam Trackings die Probanden „blind“ für die Unterschiede zwischen den verschiedenen HRTF Datenbanken machen.

6.3.3. Vergleich der HRTF Datenbanken ohne Head-Tracking

Abschließend soll beim Vergleich der HRTF Datenbanken noch das Szenario ohne Head-Tracking betrachtet werden. Die Übersicht im Boxplot findet sich in Abbildung 6.9 auf Seite 51. Der p-Wert der ANOVA beträgt $p = 0.049$, liegt also beinahe genau an der Grenze. Daher sollen auch noch die Ergebnisse des LSD Verfahrens - vgl. Abbildung 6.10 auf Seite 51 - betrachtet werden.
Der einzige signifikante Unterschied besteht demzufolge zwischen den KEMAR HRTFs und den via PLSR errechneten.
Insgesamt ist aber wohl davon auszugehen, dass auch ganz ohne den Einsatz von Head-Tracking die Wirkung der Unterschiede zwischen den einzelnen HRTF Datenbanken nicht wesentlich ins Gewicht fällt. Eine mögliche Ursache hierfür ist, dass ohne die Positionswechsel der grundsätzliche Effekt der räumlichen Wiedergabe und der Trennung der Sprecher gegenüber den Unterschieden zwischen den verschiedenen HRTFs dominiert.

6.4. Auswertung des zweiten Teils

Abbildung 6.12 auf Seite 53 zeigt die Verteilung der Beobachtungen für den zweiten Teil des Experiments, in dem der zweite Sprecher zusätzlich als Störer aktiv ist. Es zeichnet sich dabei ab, dass das vierte Treatment, mit idealer Trennung (Headset) erwartungsgemäß deutlich am besten bewertet wird. Dementsprechend iefert auch die ANOVA einen p-Wert von $p = 1.5 \cdot 10^{-12}$. Zwischen den anderen drei Treatments zeichnet sich ebenfalls eine Tendenz ab. Diese soll nun wieder mit der Least Significant DIfference Prozedur überprüft werden. Das Ergebnis ist in Abbildung 6.11 auf Seite 52 zu sehen. Ein signifikanter Unterschied besteht also nur zwischen dem Mittelwert des vierten Treatments und den drei anderen. Die Kombination aus erstem und dritten Treatment kann aber als Grenzfall ge-

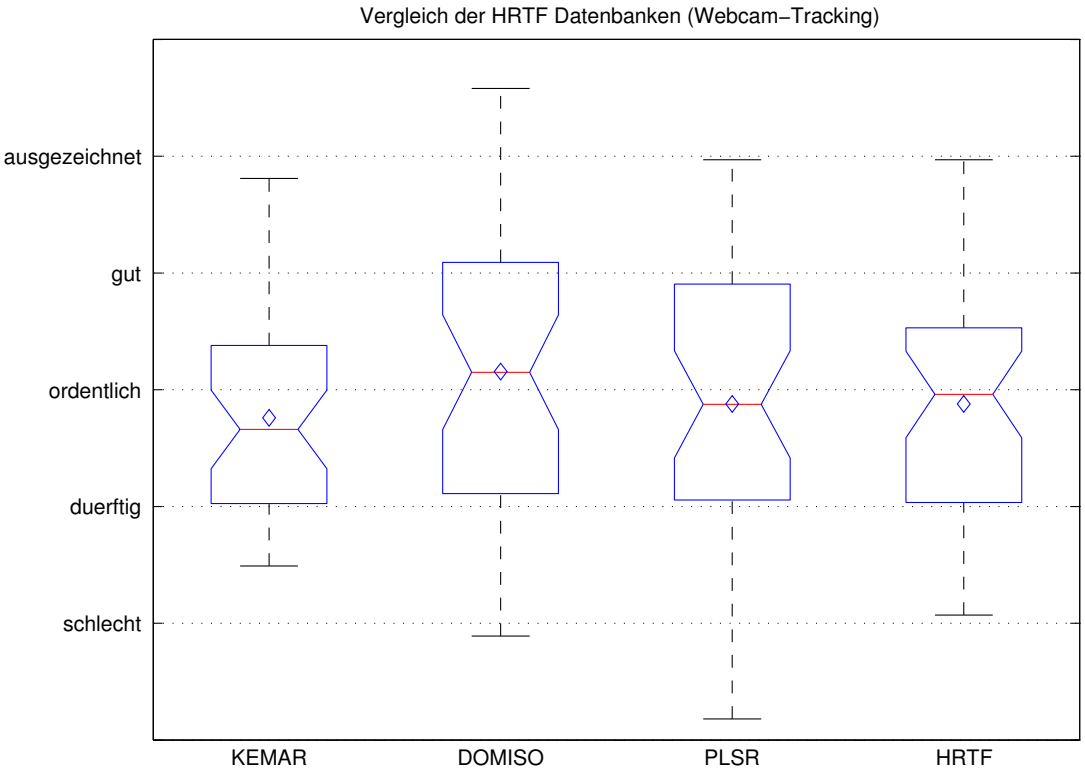


Abbildung 6.8.: Vergleich der vier HRTF Datenbanken bei Verwendung des Webcam Tracking Systems

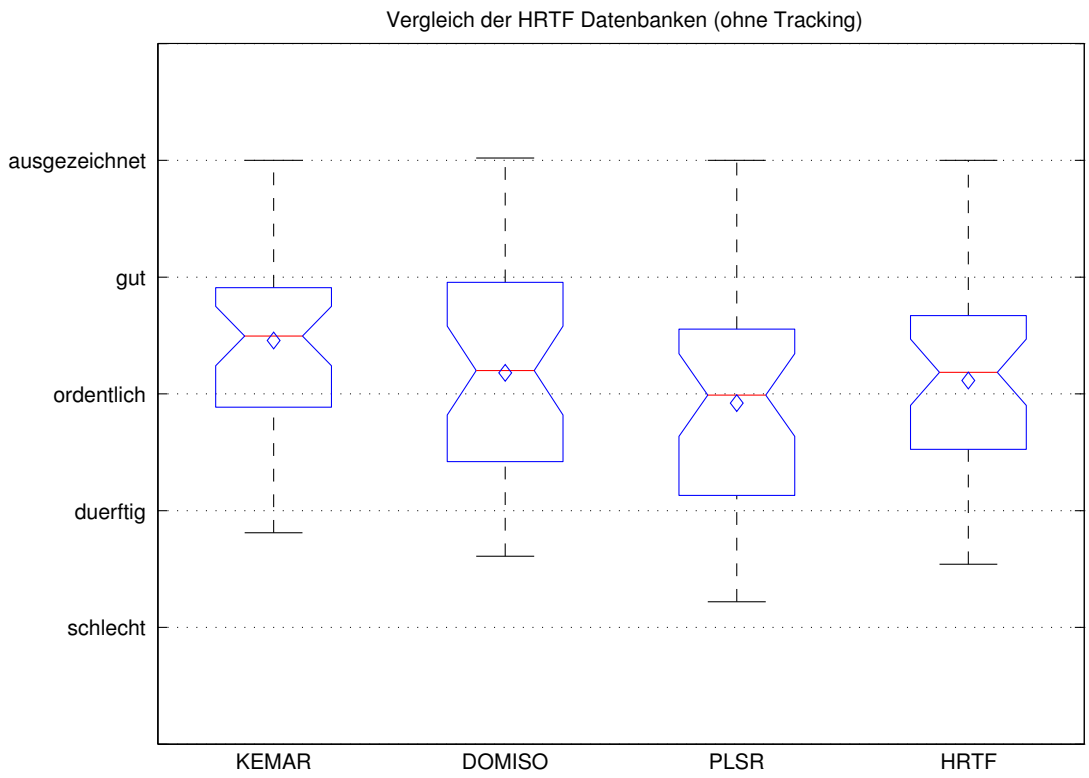


Abbildung 6.9.: Vergleich der vier HRTF Datenbanken ohne Head-Tracking

1	2	−10.14	27.75	65.63
1	3	15.74	53.63	91.51
1	4	−3.69	34.20	72.09
2	3	−12.01	25.88	63.76
2	4	−31.44	6.45	44.34
3	4	−57.31	−19.42	18.46

Abbildung 6.10.: Ergebnis des Least Siginficant Difference Verfahrens für den Vergleich der vier HRTF Datenbanken ohne Head-Tracking

1	2	−87.10	−27.25	32.61
1	3	−115.41	−55.55	4.31
1	4	−311.06	−251.20	−191.34
2	3	−88.16	−28.30	31.56
2	4	−283.81	−223.95	−164.09
3	4	−255.51	−195.65	−135.79

Abbildung 6.11.: Ergebnis des Least Siginficant Difference Verfahrens für den zweiten Versuchsteil

sehen werden, was bedeutet, dass die binaurale Wiedergabe auch bei starken Artefakten aus dem Channel Assignement tendenziell bevorzugt wird.

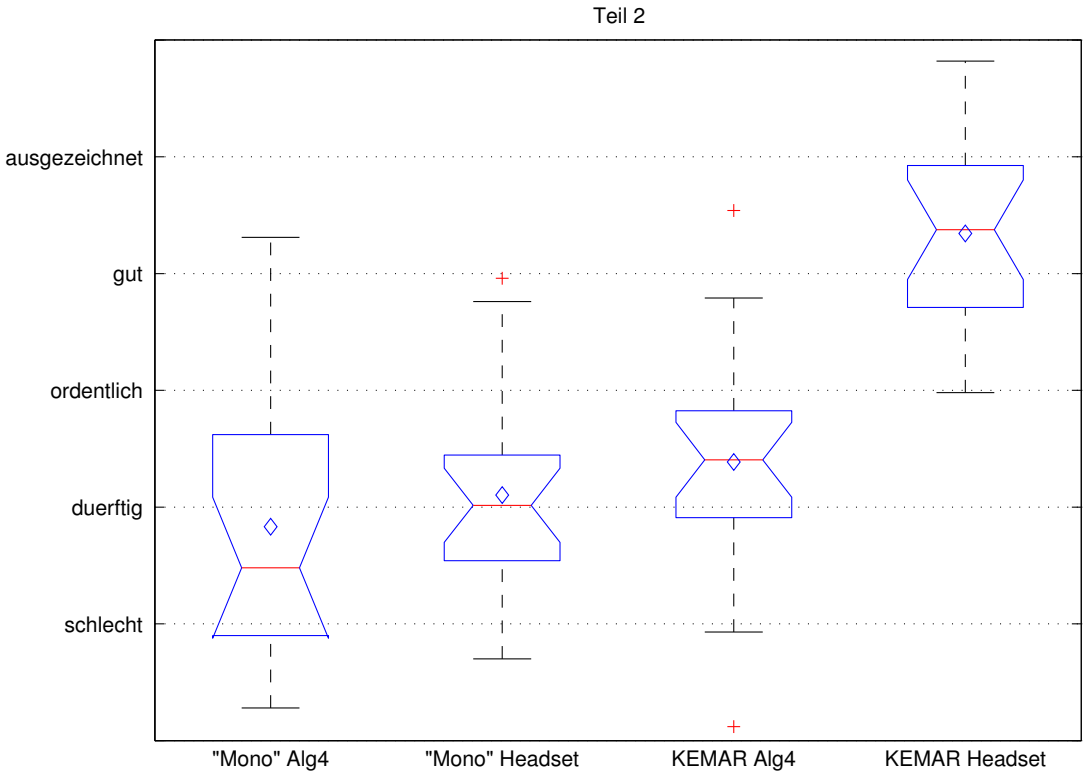


Abbildung 6.12.: Boxplot der Ergebnisse des zweiten Versuchsteils

7. Resumee und Ausblick

Es gelang im Rahmen dieser Arbeit, ein umfangreiches Experiment zur QoE Evaluierung des Telekonferenzsystems im Entwicklungsstadium durchzuführen. Dazu wurde eine Analyse des QoE Begriffs im Gesamtkontext möglicher Qualitätsevaluierungsansätze durchgeführt und aufbauend auf einer strukturellen Analyse des Systems und einer Literatur Recherche zur Planung und Durchführung von Versuchen zur subjektiven Evaluierung ein eigenes Konzept zur QoE Evaluierung erarbeitet.
Die Ergebnisse des Experiments geben Aufschluss über die systematische Wirkung der verschiedenen Optionen für das Channel Asignement, das Head-Tracking und die binaurale Wiedergabe auf die vom Nutzer wahrgenommene QoE. Die wichtigsten Erkenntnisse aus dem Experiment sollen im Folgenden noch einmal kurz zusammengefasst werden.
Der am LDV implementierte Algorithmus 4 zum Channel Assignement schneidet bei der Evaluierung zwar etwas schlechter ab als eine durch separate Mikrofone (ideal) getrennte Aufnahme, ist aber mit einer mittleren Bewertung „ordentlich“ gegenüber einem knappen „gut“ für das Headset Szenario durchaus konkurrenzfähig.
Beim Vergleich der Head-Tracking Optionen ergab sich erwartungsgemäß ein klarer Vorteil für das AR Tracking System mit seinen drei Infrarotkameras, dass aber in erster Linie als Referenz im Versuch eingesetzt wurde und für ein fertiges Produkt keine denkbare Lösung ist. Etwas überraschend war das schwache Abschneiden des am LDV implementierten Webcam Trackingverfahrens, das sogar schlechter als die Variante ganz ohne Head-Tracking bewertet wurde. Daraus lässt sich schließen, dass es sich bei der weiteren Arbeit an dem System lohnen würde, an einem leistungsstärkeren, produkt-tauglichen Trackingverfahren zu arbeiten.
Bei der Evaluierung der HRTF Datenbanken ergab sich auf den ersten Blick der Trend, dass die Unterschiede zwischen den einzelnen HRTF Datenbanken für die Teilnehmer nur schwer wahrnehmbar waren. Lediglich die via PLSR anhand anthropometrischer Daten errechneten HRTFs wurden allgemein schlechter bewertet. Eine genauere Analyse der Daten unter Berücksichtigung der unterschiedlichen Tracking Optionen ergab jedoch, dass bei Verwendung des AR Trackings ein Vorteil für die individuellen HRTFs gegenüber den aus PLSR und DOMISO Selection ermittelten HRTFs zu erkennen ist. Dieses Ergebnis zeigt, dass der in der Literatur weithin bekannte positive Einfluss von Head-Tracking [12] sich auch in Form einer genaueren Wahrnehmung der Unterschiede zwischen einzelnen HRTF Datenbanken abzeichnet.
Für zukünftige Arbeiten wäre insbesondere eine genauere Evaluierung der Wirkung der verschiedenen HRTF Datensätze in einem separaten Experiment von großem Interesse. Aus der Literatur [11] geht hervor, dass bei Verwendung unterschiedlicher HRTFs unter-

schiedliche Resultate bei Lokalisierungstests erzielt werden, d.h. es besteht grundsätzlich Anlass zu der Vermutung, dass es einen solchen Unterschied gibt. Es ist durchaus denkbar, dass dieser im Experiment der vorliegenden Arbeit aufgrund anderer, stärker wahrgenommener Effekte nur schwach ausgeprägt zu beobachten war oder allgemein nur in Experimenten mit größerer Expertise der Probanden zu Tage tritt.

Literaturverzeichnis

1. *ITU-T P.800 Methods for Subjective Determination of Transmission Quality*. International Telecommunication Union, 1996.

2. *ITU-R BS.1116-1 Methods for the Subjective Assessment of Small Impairments in Audio Systems Including Multichannel Sound Systems*. International Telecommunication Union, 1997.

3. *ITU-R BS.1284-1 General Methods for the Subjective Assessement of Sound Quality*. International Telecommunication Union, 2003.

4. *ITU-R BS.1534-1 Method for the Subjective Assessment of Intermediate Quality Level of Coding Systems*. International Telecommunication Union, 2003.

5. *ITU-T E.800 Definitions of Terms Relataed to Quality of Service*. International Telecommunication Union, 2008.

6. *ITU-T P.10/G.100 Vocabulary for Performance and Quality of Service (Amendment 2)*. International Telecommunication Union, 2008.

7. *DTrack2 User Manual*. Advanced Realtime Tracking GmbH, 2011.

8. *ITU-T P.1301 Subjective Quality Evaluation of Audio and Audiovisual Multiparty Telemeetings*. International Telecommunication Union, 2012.

9. V.R. Algazi, R.O. Duda, D.M. Thompson und C. Avendano. The cipic hrtf database. In *IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*. 2001.

10. S. Bech und N. Zacharov. *Perceptual Audio Evaluation - Theory, Method and Application*. John Wiley & Sons Ltd, 2007.

11. D. Begault, E. Wenzel und M. Anderson. Direct comparison of the impact of head tracking, reverberation and individualized head-related transfer functions on the spatial perception of a virtual speech source. In *Journal of the Audio Engineering Society*, 2001.

12. D.R. Begault. Virtual acoustic displays for teleconferencing: Intelligibility advantage for 'telephone-grade' audio. In *Journal of the Audio Engineering Society*, 1999.

13. J. Berg und F. Rumsey. Systematic evaluation of perceived spatial quality. In *Audio Engineering Society Conference: 24th International Conference: Multichannel Audio, The New Reality*. Audio Engineering Society, 2003.

14. J. Blauert. *Spatial Hearing*. The MIT Press, 1999.

15. K. Blum, G.J. Van Rooyen und H.A. Engelbrecht. Spatial audio to assist speaker identification in telephony. In *17th International Conference on Systems, Signals and Image Processing*, 2011.

16. M. Bodden und U. Jekosch. *Entwicklung und Dürchführung von Tests mit Versuchspersonen zur Verifizierung von Modellen zur Berechnung der Sprachübertragungsqualität*. unveröffentlicht, 1996.

17. D.S. Brungart und B.D. Simpson. *Improving Multitalker Speech Communication With Advanced Audio Displays*. Technischer Bericht, DTIC Document, 2005.

18. P. Burger und M. Rothbucher. *Self Initializing Head Pose Estimation with a 2D Monocular USB Camera*. Technischer Bericht, Technische Universität München, 2013.

19. P.L. Callet, S. Möller und A. Perkis. Qualinet white paper on definitions of quality of experience. 2012.

20. E.Y. Choueiri. *Optimal Crosstalk Cancellation for Binaural Audio with Two Loudspeakers*. Technischer Bericht, Princeton University, 2008.

21. R. Drullman und A.W. Bronkhorst. Multichannel speech intelligibility and talker recognition using monaural, binaural, and three-dimensional auditory presentation. In *The Journal of the Acoustical Society of America*, 2000.

22. H. Fastl. The psychoacoustics of sound-quality evaluation. In *Acta Acustica united with Acustica*, 1997.

23. H. Fastl und E. Zwicker. *Psychoacoustics: Facts and Models*. Springer, 2007.

24. A. Field. *Discovering statistics using SPSS*. Sage publications, 2009.

25. B. Gardner und K. Martin. *HRTF Measurements of a KEMAR Dummy-Head Microphone*. Technischer Bericht, MIT Media Lab, 1994.

26. T. Grasser. *Auswahlverfahren für HRTFs zur 3D Sound Synthese*. Masterarbeit, Technische Universität München, 2012.

27. T. Grasser. *Speaker Localization and Separation in Teleconferences*. Masterarbeit, 2013.

28. N. Inoue, T. Kimura, T. Nishino, K. Itou und K. Takeda. Evaluation of hrtfs estimated using physical features. In *Acoustical science and technology*, 2005.

29. Y. Iwaya. Individualization of head-related transfer functions with tournament-style listening test: Listening with other's ears. In *Acoustical Science and Technology*, 2006.

30. D.L. Jones, K.M. Stanney und H. Foaud. An optimized spatial audio system for virtual training simulations: Design and evaluation. In *International Conference on Auditory Display*. 2005.

31. A. Kuhn. *HRTF Customization by Regression*. Masterarbeit, Technische Universität München, 2013.

32. T. Letowski. Sound quality assessment: Concepts and criteria. In *Audio Engineering Society Convention 87*. Audio Engineering Society, 1989.

33. A. Lindau und S. Weinzierl. On the spatial resolution of virtual acoustic environments for head movements in horizontal, vertical and lateral direction. In *EAA Symposium on Auralization*. 2009.

34. G. Lorho. *Percieved Quality Evaluation - An Application to Sound Reproduction over Headphones*. Dissertation, Aalto University School of Science and Technology, 2010.

35. P. Mackensen. *Auditive Localization. Head Movements, an Additional Cue in Localization*. Dissertation, Technische Universität Berln, 2004.

36. H. Martens und M. Martens. *Multivariate Analysis of Quality - An Introduction*. John Wiley & Sons Ltd, 2001.

37. S. Möller. *Quality of Telephone-Based Spoken Dialogue Systems*. Springer, 2004.

38. S. Möller. *Quality Engineering*. Springer, 2010.

39. T. Neher, F.J. Rumsey und T. Brookes. Training of listeners for the evaluation of spatial sound reproduction. In *Audio Engineering Society Convention 112*. Audio Engineering Society, 2002.

40. A. Raake. *Speech Quality of VoIP - Assessment and Prediction*. John Wiley & Sons Ltd, 2006.

41. M. Rothbucher, M. Kaufmann, T. Habigt, J. Feldmaier und K. Diepold. Backwards compatible 3d audio conference server using hrtf synthesis and sip. In *Seventh International Conference on Signal-Image Technology and Internet-Based Systems (SITIS)*. 2011.

42. M. Rothbucher, K. Veprek, P. Paukner, T. Habigt und K. Diepold. Comparison of head-related impulse response measurement approaches. In *Journal of the Acoustical Society of America Express Letters*, 2013.

43. F. Rumsey. Spatial quality evaluation for reproduced sound: Terminology, meaning, and a scene-based paradigm. In *Journal of the Audio Engineering Society*, 2002.

44. F. Rumsey, S. Bech et al.. On some biases encountered in modern audio quality listening tests - a review. In *Journal of the Audio Engineering Society*, 2008.

45. J. Sandvad. Dynamic aspects of auditory virtual environments. In *Audio Engineering Society Convention 100*. Audio Engineering Society, 1996.

46. B. Seeber, H. Fastl et al.. Subjective selection of non-individual head-related transfer functions. In *Proceedings of the International Conference on Auditory Display*. 2003.

47. J. Skowronek, A. Raake, K. Hoeldtke und M. Geier. Speech recordings for systematic assessment of multi-party conferencing. In *Proceedings of Forum Acusticum*. 2011.

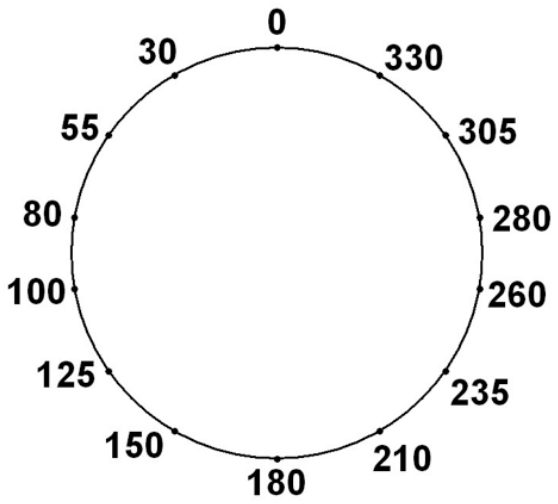
48. K. Steierer. *Teleconference Channel Assignement*. Masterarbeit, 2013.

49. T. Volk. *Planung, Einrichtung und akustische Vermessung eines reflexionsarmen Raumes zur Untersuchung von HRTFs*. Masterarbeit, Technische Universität München, 2011.

50. E.M. Wenzel. What perception implies about implementation of interactive virtual acoustic environments. In *Audio Engineering Society Convention 101*. Audio Engineering Society, 1996.

A. Einführungsblatt zur HRTF Selection (DOMISO)

Der folgende Hörversuch ist die Implementierung eines „HRTF-Selection-Verfahrens“. Neben der Vermessung der eigenen HRTFs, die wir ja bereits mit ihnen durchgeführt haben, gibt es auch die Möglichkeit, die HRTFs einer anderen Person zur Synthese eines dreidimensionalen Schallfeldes zu verwenden.
Dieser Test dient dazu, aus einer Datenbank, die aus den HRTFs zwölf fremder Personen besteht, denjenigen Datensatz auszuwählen, der am besten zu ihnen passt, d.h. ihnen den besten räumlichen Klangeindruck ermöglicht.
Im Verlauf des Versuchs werden ihnen Testgeräusche vorgespielt in denen sogenanntes „Rosa-Rauschen“ eine kreisförmige Bewegung entgegen dem Uhrzeigersinn um ihren Kopf herum durchläuft.
Diese Testgeräusche werden in paarweisen Vergleichen präsentiert, bei denen sie sich dann für den jeweils besseren Kandidaten entscheiden sollen. Die Testgeräusche können bei Bedarf auch mehrmals angehört werden.
Zur Entscheidung können sie neben ihrem allgemeinen Gesamteindruck auch darauf achten ob sie das Geräusch tatsächlich „außerhalb“ ihres Kopfes wahrnehmen und wie gut sie die Kreisbewegung entlang der vorgegebenen Positionen (s. Abbildung unten) nachvollziehen können.
Falls sie noch Fragen zum Ablauf des Tests haben können sie diese gerne noch vor Beginn stellen.



B. Einführungsblatt zum Hörversuch

Hörversuch „Telekonferenz“

Der folgende Hörversuch dient der Evaluierung eines Telekonferenzsystems. Die Besonderheit dieses Systems ist die räumlich getrennte Wiedergabe der einzelnen Konferenzteilnehmer auf einem Stereo-Kopfhörer.

Im Laufe des Versuchs werden Sie einen ca. 60 Sekunden langen Ausschnitt aus einer Telekonferenz mehrmals vorgespielt bekommen. Dabei werden bei jedem Durchgang verschiedene Wiedergabeverfahren und –algorithmen zum Einsatz kommen.
Nach jedem Durchgang bitten wir Sie, das jeweils gerade gehörte Beispiel hinsichtlich seiner Qualität zu bewerten. Sie können dazu ihre eigenen Kriterien verwenden. Zusätzlich werden wir sie vor Versuchsbeginn im Rahmen eines kurzen Trainings auf einige mögliche Kriterien zur Bewertung aufmerksam machen.

Die Bewertung wird jeweils auf einer Skala vorgenommen, wie sie auf der ersten Abbildung unten zu sehen ist. **Die Skala ist kontinuierlich** und auch die Randbereiche können verwendet werden.

Im Verlauf der Konferenz werden vier Sprecher zu Wort kommen. Jeder dieser Sprecher sollte an seiner vorgegebenen Position zu hören sein (s. zweite Abbildung unten).
Wir bitten Sie bei Ihrer Beurteilung sowohl Gesprächspassagen in denen nur ein Sprecher spricht, als auch schnelle Wortwechsel und Unterbrechungen in Betracht zu ziehen. Wir bitten sie außerdem, die **Bewertung zügig und intuitiv** vorzunehmen. **Dieser Versuch ist rein subjektiver Natur, daher gibt es weder richtige noch falsche Antworten.**
Wenn sie Fragen zum Ablauf des Versuchs haben können sie diese gerne vor Beginn noch stellen.

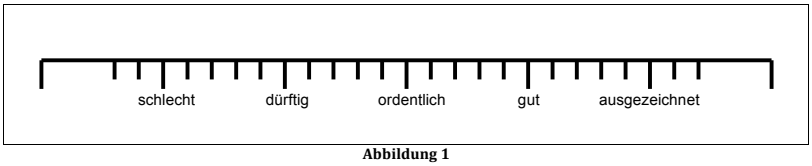


Abbildung 1

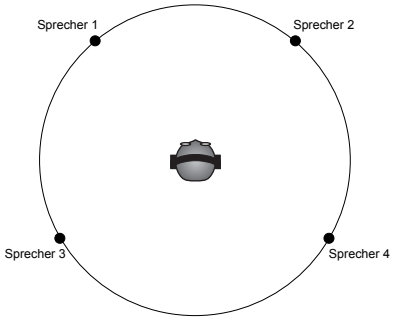


Abbildung 2

C. Einführungsblatt zum Training

Training zum Hörversuch „Telekonferenz“

Im Laufe des Trainings zum folgenden Hörversuch werden Sie insgesamt sechs Audiobeispiele zu hören bekommen. Im Folgenden finden Sie jeweils eine kurze Erläuterung zu den einzelnen Beispielen.
In allen Trainingsbeispielen hören Sie, wie sich die 4 Teilnehmer zu Beginn der Konferenz kurz vorstellen.

Beispiel 1:
In diesem Beispiel kommt, wie bei den meisten klassischen Telekonferenzsystemen üblich, eine einfache Mono-Wiedergabe zum Einsatz.

Beispiel 2:
In diesem Beispiel sind die Sprecher nun räumlich getrennt zu hören.

Beispiel 3:
Nun wird zusätzlich zur räumlich getrennten Wiedergabe die Bewegung ihres Kopfes von unserem System erfasst. Wenn Sie nun den Kopf hin und her bewegen, sollte das Klangbild stabil „stehen“ bleiben.

Beispiel 4:
Nun hören Sie noch einmal Beispiel 2 um den Unterschied zu Beispiel 3 zu verdeutlichen.

Beispiel 5:
Sie hören noch einmal Beispiel 3 um Sie an das Headtracking (erfassen der Kopfbewegung durch das System zu gewöhnen). Bewegen Sie gerne den Kopf wieder, um die Auswirkungen des Trackings gut zu hören.

Beispiel 6:
In diesem Beispiel werden Sie einige Artefakte, die bei der Datenverarbeitung in dem Telekonferenzsystem entstehen können zu hören bekommen. Achten Sie dabei besonders auf den Wechsel zwischen dem ersten und dem zweiten Sprecher (leichte Artefakte), sowie auf den Übergang zwischen Sprecher 3 und 4 (starke Artefakte).

D. Boxplot: Übersicht über alle 28 Treatments

Die Abkürzungen an der x-Achse sind wie folgt zu verstehen:

KEMAR/DOMISO/PLSR/HRTF/MONO beschreibt die Variante der binauralen Wiedergabe, wobei HRTF den individuellen Datensatz und MONO die Wiedergabe mit KEMAR HRTFs und allen vier Sprechern auf der selben Position meint.

AR/FT/NT gibt Auskunft über das verwendete Head-Tracking, wobei FT für das USB Tracking steht und NT für die „Off-Position“.

A4/HS gibt schließlich die Variante des Channel Assignement an, wobei A4 für Algorithmus 4 und HS für das headset Szenario steht.

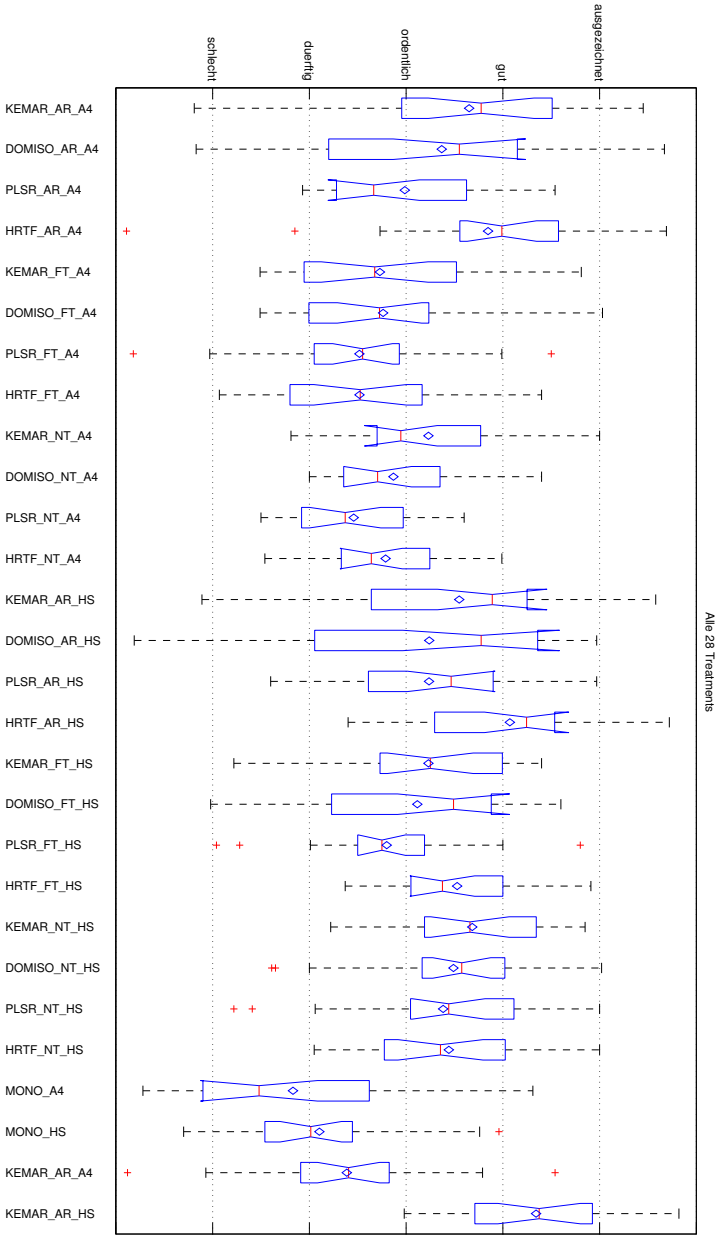


Abbildung D.1.: Boxplot-Übersicht über alle 28 Treatments