Perzeptive Audio-Qualitätsdimensionen im audio-visuellen Kontext

Falk Schiffner[(a)] & Sebastian Möller [(a,b)]

[(a)] Quality and Usability Lab, Technische Universität Berlin, Germany

[(b)] German Research Center for Artificial Intelligence (DFKI)

Email: [falk.schiffner, sebastian.moeller]@tu-berlin.de

Einleitung

Audio-visuelle-Kommunikationsdienste sind weit verbreitet und der Alltag ist im Allgemeinen komplett mit deren Nutzung durchdrungen. Anbieter dieser Dienste betreiben großen Aufwand um die „Qualität“ auf hohem Niveau zu liefern. Die Schätzung der wahrgenommenen audio-visuellen Qualität dieser Dienste geschieht oftmals mittels instrumenteller Verfahren (ITU-T G.1070 [1]). Diese Schätzer liefern meist nur einen MOS (Mean Opinion Score) und keine weiteren diagnostischen Informationen. Hier kann der Ansatz der Schätzung mittels wahrnehmungsbasierten Modellen verwendet werden. Die Grundidee dahinter besteht darin, dass ein „Ereignis“ in einem kognitiven Wahrnehmungsraum stattfindet. Um nun dieses Ereignis zu quantifizieren, werden die relevanten Eigenschaften intern bewertet, d.h. auf den jeweiligen Raumachsen positioniert. Liegen die Achsen rechtwinklig zueinander, können die Wahrnehmungseigenschaften als Wahrnehmungsdimensionen erachtet werden. Der Ansatz, die Qualitätsschätzung mittels eines dimensions-basierten Modells zu verbessern, wurde bereits bei der Sprachtelefonie verfolgt [2], [3].

In diesem Beitrag wird untersucht, ob dieser Ansatz auch für den audio-visuelle Kontext (bspw. Videotelefonie) verwendet werden kann. Es wurden bereits Untersuchungen für den Videokanal durchgeführt und der Wahrnehmungsraum aufgestellt [4]. Die qualitätsrelevante Wahrnehmungsräume für Video- und Sprachsignalübertragung bestehen aus 5 Dimensionen für Video und 4 Dimensionen für Sprache. Eine kurze Beschreibung der Dimensionen ist in Tabelle 1 zu finden.

Froschungsfragen

Wie in [5] beschrieben, wird die Audio- und Videoqualität getrennt von einander beurteilt und erst im nachfolgenden Schritt zur Gesamtqualität zusammengefügt. Auf dieser Grundlage stellen sich die Fragen, mit den sich dieser Beitrag befasst:

Tabelle 1: Beschreibung der perzeptiven Qualitätsdimensionen für Video (VQD) und Audio (AQD).

VQD	Name	Beschreibung
I	*Fragmentierung (FRA)*	Video ist zerfallen, unzusammenhängend
II	*Unklarheit (UCL)*	Unklares und verschwommenes Video
III	*Diskontinuität$_{video}$ (DIC_v)*	Unterbrechungen im Fluss des Videos
IV	*Rauschhaftigkeit$_{video}$ (NOI_v)*	Zufällige Änderungen d. Helligkeit / Pel
V	*Suboptimale Helligkeit (LUM)*	Zu hohe oder zu geringe Helligkeit

AQD	Name	Beschreibung
I	*Klangverfärbung (COL)*	Unnatürlicher, metallischer Klang
II	*Diskontinuität$_{audio}$ (DIC_a)*	Unterbrechungen im Fluss im Sprachsig.
III	*Rauschhaftigkeit$_{audio}$ (NOI_a)*	Unerwünschte Rauschgeräusche
IV	*Lautheit (LOU)*	Zu lautes / zu leises Sprachsignal

- *Können die perzeptiven Qualitätsdimensionen aus der Sprachtelefonie auch im Videotelefonie-Kontext verwendet werden?*
- *Unterscheiden sich die Bewertungen der Audiodimensionen, wenn der Videokanal aktiv ist?*
- *Unterscheiden sich die Bewertungen der Audiodimensionen bei unterschiedlichen Videostörungen?*

Um diese Fragen zu beantworten, wurde ein Experiment durchgeführt, bei dem Audio- zusammen mit Videostörungen untersucht wurden. Der Beitrag präsentiert sowohl die Versuchsdurchführung, also auch die Ergebnisse und Schlussfolgerungen.

Experiment

Versuchsteilnehmer: Für diese Studie haben wir 32 Probanden eingeladen. Die Teilnehmergruppe bestand aus 17 weiblichen und 15 männlichen Teilnehmer, welche über ein Probandenprotal akquiriert wurden. Das durchschnittliche Alter betrug $32,5$ Jahre ($\sigma = 9,3$). Bei jedem Teilnehmer wurde vor den Versuch ein Sehtest durchgeführt (*Ishihara-Test* [6], *Snellen Table* [7]). Es musste kein Teilnehmer aufgrund einer Sehstörung ausgeschlossen werden. Es wurde kein Hörscreening durchgeführt. Alle Teilnehmer berichteten, dass ihnen kein Hörbeeinträchtigung bekannt sei. Den Teilnehmern war es erlaubt die Lautstärke auf ihren individuelle MCL (*Most Comfort Level*) einstellen. Es kann davon ausgegangen werden, dass die Lautstärke ausreichend hoch war. Alle Teilnehmer haben eine Vergütung für die Testteilnahme erhalten.

Testprozedur: Die Testdauer betrug etwa $55\,min$. Die Teilnehmer konnte auf Wunsch eine Pause von $5\,min$ nehmen, falls sie diese benötigen. Damit sollte ein eventuelle Ermüdung verhindert werden. Zu Beginn wurde eine schriftliche Testeinweisung gegeben. Weiterhin wurde ein kurzes Training durchgeführt, damit sich die Teilnehmer mit der Testaufgabe und den Störungen vertraut machen konnten. Die Reihenfolge der Testdateien war für jeden Teilnehmer randomisiert, um potenzielle Reihenfolgeeffekte auszuschließen. Die Teilnehmer mussten jede Testdatei einmal komplett abspielen und konnte es anschließend so oft wiederholen wie sie wollten. Die Bewertungsaufgabe war in drei Teile untergliedert. Im ersten Teil, mussten die Teilnehmer die Gesamtqualität, die Video- und Audioqualität der Probe auf einer 7-stufigen kontinuierlichen Skala (Abbildung 1) [8] bewerten. Im zweiten und dritten Teil, mussten die Teilnehmer bewerten, in welchen Maße die qualitätsrelevanten Dimensionen für Audio und Video in der Probe vorhan-

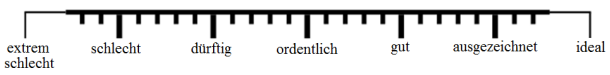

Abbildung 1: 7-stufige kontinuierliche Skala, verwendet für die Gesamt-, Audio und Videoqualitätsbewertung, (korrespondierte Skalenwerte in Klammern): extrem schlecht (10), schlecht (20), dürftig (30), ordentlich (40), gut (50), ausgezeichnet (60) und ideal (70).

den war. Die Reihenfolge von Teil zwei und drei wurde zwischen den Testteilnehmern gewechselt. Zusätzlich wurde für jeden Teilnehmer die Reihenfolge der Bewertungsskala randomisiert. Die Dimensionsnamen wurden als Skalenüberschrift verwendet und entsprechende Antonympaare wurden verwendet, um die Skalenweite zubeschreiben (basierend auf Likert-Skala [9]; vgl. Abbildung 2). Die rechte Seite der jeweiligen Skala kann als optimal für die entsprechende Eigenschaft erachtet werden. Die Skalenweite der für die Dimensionsbewertung und Qualitätsbewertung ist identisch und reicht von $10-70$. Wobei 70 dem Optimum für die jeweilige Eigenschaft entspricht. Der Test wurde in einem Versuchsraum nach ITU-T Rec. P.910 [10] durchgeführt.

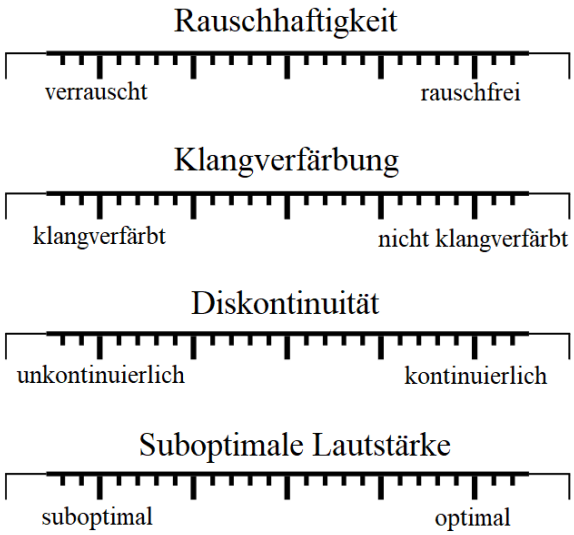

Abbildung 2: Bewertungsskalen für die Audio-Qualitätsdimensionen, Dimensionsnamen als Skalenüberschrift mit Antonympaar als Skalenbeschriftung.

Versuchsmaterial und Beeinträchtigungen: Das Ziel bei der Wahl der Beeinträchtigungen war es, eine möglichst große Spanne von Übertragungsstörungen abzubilden. Zusätzlich sollte auch die Vergleichbarkeit mit vorherigen Studien gewährleistet sein, sodass Ankerkonditionen eingefügt wurden. Da sich der Beitrag auf die Audioqualität und die Audioqualitätsdimensionen konzentriert, werden die Beeinträchtigungen im Video und die Bewertungen der Videoqualitätsdimensonen hier nicht betrachtet.

Genauso wie beim Videokanal, wurde der Audiokanal mit Störungen behaftet, die üblich für die Video- und Sprachsignalübertragung sind. Um den Audiokanal zu beeinträchtigen, haben wir die *Software Tool Library* verwendet, die von der ITU-T bereit gestellt wird [11]. Wir haben drei „Störungsklassen“ erstellt. Erstens, nur der Audiokanal wurde gestört und zusammen mit dem Referenzvideo abgespielt. Zweitens, nur der Videokanal wurde gestört und zusammen mit dem Referenzaudio abgespielt. Drittens, beide Signalkanäle wurden in unterschiedlichen Kombinationen gestört. Es wurde bedacht, dass alle perzeptiven Dimensionen durch die Signalstörungen stimuliert werden. Außerdem wurde jede Kombination einer Video- mit einer Audiodimension im Test untersucht. In der Tabelle 2 sind die verschiedenen Beeinträchtigungen und die korrespondierenden Qualitätdimensionen aufgelistet. In Abbildung 3 ist beispielhaft ein Screenshot zweier Testdateien zu sehen.

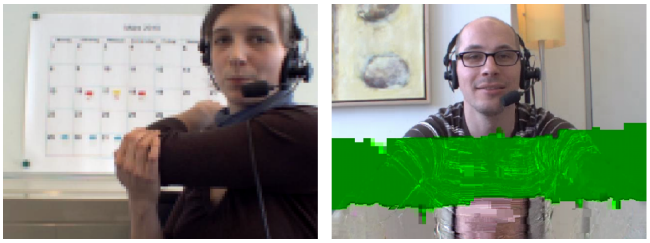

Abbildung 3: Beispiel – Videomaterial, Kopf-und-Schulter Szene, links: Blockeffekt $5x5pel$, rechts: Packetverlust $1,5\%$.

Ergebnisse

Die Qualitätsbewertungen und die Bewertungen der einzelnen Dimensionen wurden ausgewertet. Wir transformierten die Bewertungen von der 7-stufigen Skala zu einer 5-stufigen MOS (Mean Opinion Score) Skala mittels der Formel aus [12]. Im Allgemeinen, kann beobachtet werden, dass die Beeinträchtigungen auf die *„richtigen“* Wahrnehmungsdimensionen anspringen. Die Ergebnisse sind in Tabelle 2 gegeben.

Beim Blick auf die Qualitätsbewertung, ergibt sich, wie erwartet, dass je stärker die Beeinträchtigung ist, desto geringer ist die Qualitätsbewertung (bspw. Tabelle 2 Kondition 03 und 04). Weiterhin zeigt sich, dass unterschiedliche Störungen im Videokanal sich nicht auf die Bewertung der Sprachqualitätsdimensionen auswirken. Die mittlere Standardabweichung für die Bewertung auf den Sprachdimensionsskalen beträgt nur $0,24$, also weniger als ein viertel MOS. Die mittlere Standardabweichung für die Bewertung der Gesamtsprachqualität ist ebenfalls gering und beträgt $0,29$. Aus den Ergebnissen lässt sich schließen, dass das Bewertungsverhalten auf den Sprachdimensionsskalen, gleich ist und somit unabhänig vom Videokanal und den dort eingebrachten Störungen.

Qualitätsmodellierung: Um die Gesamtqualität des Sprachkanals aus den Dimensionsbewertungen zu modellieren, verwenden wir das lineare Modell aus [13]. Das Modell wurde in einem reinen Audio-Kontext erstellt. Die Sprachgesamtqualität wird mittels der nachstehenden Formel berechnet.

Tabelle 2: Transformierte Qualitätsbewertungen – Q (Gesamtqualität), VQ (Videoqualität), AQ (Audioqualität) und die transformierten Bewertungen der perzeptiven Audioqualitätsdimensionen (DIM_a).

#	Video Impairment	Audio Impairment	DIM_v	DIM_a	Q	VQ	AQ	COL_a	NOI_a	DIC_a	LOU_a
00	Reference	Reference	–	–	4,18	4,30	3,96	4,33	3,73	4,47	4,34
01	Reference	Level $-16dBov$	–	LOU	3,40	4,19	2,64	4,03	3,80	4,43	1,84
02	Reference	Level $-36dBov$	–	LOU	4,17	4,27	4,03	4,20	4,24	4,36	4,14
03	Reference	G.722.2 $PL0.2\%$	–	DIC_a	3,81	4,13	3,37	3,78	4,05	3,64	4,13
04	Reference	G.722.2 $PL0.8\%$	–	DIC_a	3,06	4,08	2,03	3,33	3,59	2,01	3,95
05	Reference	G.711	–	COL	3,05	4,08	2,03	2,34	2,33	3,99	3,53
06	Reference	G.722.2 $(6.6kbits)$	–	COL	3,80	4,12	3,53	3,51	3,94	4,32	4,16
07	Reference	MNRU-45	–	NOI_a	4,12	4,32	4,02	4,27	4,09	4,41	4,11
08	Reference	MNRU-55	–	NOI_a	4,12	4,34	3,82	4,13	4,02	4,44	3,78
09	Luminance (dark)	Reference	LUM	–	2,83	2,18	4,00	4,18	4,08	4,45	4,10
10	Luminance (light)	Reference	LUM	–	3,21	2,70	3,90	4,06	4,17	4,40	4,04
11	Blurring (Filter 1)	Reference	UCL	–	2,90	2,38	3,79	3,81	3,93	4,30	4,09
12	Jerkiness (6 Frames)	Reference	DIC_v	–	2,93	2,50	3,67	3,76	3,93	3,96	3,99
13	NoiseQ 3%	Reference	NOI_v	–	2,73	2,10	3,99	4,16	4,18	4,42	4,23
14	Packet Loss 1.5%	Reference	FRA	–	2,34	2,00	2,87	3,12	3,64	3,58	3,90
15	Luminance (dark)	Level $-36dBov$	LUM	LOU	2,91	2,22	3,86	4,17	4,28	4,38	4,19
16	Luminance (dark)	G.722.2 $PL0.8\%$	LUM	DIC_a	2,33	1,97	2,99	3,38	3,31	3,33	3,84
17	Luminance (light)	G.722.2 $(6.6kbits)$	LUM	COL	2,84	2,43	3,26	3,30	3,62	4,17	3,90
18	Luminance (light)	MNRU-45	LUM	NOI_a	3,13	2,50	3,83	4,05	4,00	4,42	4,16
19	Blurring (Filter 1)	Level $-36dBov$	UCL	LOU	2,84	2,38	3,80	3,95	3,99	4,37	4,02
20	Blurring (Filter 1)	G.722.2 $PL0.8\%$	UCL	DIC_a	2,50	2,22	2,76	3,39	3,59	2,87	3,98
21	Blurring (Filter 1)	MNRU-45	UCL	NOI_a	2,60	2,12	3,53	3,88	3,78	4,27	3,75
22	Blurring (Filter 1)	G.711	UCL	COL	2,09	2,04	2,15	2,44	2,58	3,92	3,71
23	Jerkiness (6 Frames)	Level $-16dBov$	DIC_v	LOU	2,55	2,38	2,73	3,77	3,51	4,15	1,81
24	Jerkiness (6 Frames)	G.722.2 $PL0.8\%$	DIC_v	DIC_a	2,46	2,52	2,38	3,09	3,70	2,56	3,87
25	Jerkiness (6 Frames)	G.722.2 $(6.6kbits)$	DIC_v	COL	2,59	2,54	2,57	2,88	3,46	3,07	3,83
26	Jerkiness (6 Frames)	MNRU-45	DIC_v	NOI_a	2,90	2,53	3,54	3,83	3,73	3,90	4,01
27	NoiseQ 3%	Level $-16dBov$	NOI_v	LOU	2,31	2,03	2,61	3,79	3,44	4,16	2,01
28	NoiseQ 3%	G.722.2 $PL0.8\%$	NOI_v	DIC_a	2,16	1,97	2,41	3,26	3,48	2,75	3,96
29	NoiseQ 3%	G.722.2 $(6.6kbits)$	NOI_v	COL	2,39	1,94	3,05	3,05	3,45	3,80	3,80
30	NoiseQ 3%	MNRU-45	NOI_v	NOI_a	2,54	2,03	3,49	3,90	3,45	4,30	4,00
31	Packet Loss 1,5%	Level $-16dBov$	FRA	LOU	1,98	1,83	2,20	3,06	3,39	3,70	1,71
32	Packet Loss 1,5%	G.722.2 $(6.6kbits)$	FRA	COL	1,97	1,81	1,92	2,69	3,25	1,98	3,56
33	Packet Loss 1,5%	MNRU-45	FRA	NOI_a	2,20	1,85	2,60	2,97	3,60	3,22	3,92
34	Packet Loss 1,5%	G.711	FRA	COL	1,73	1,76	1,79	2,40	2,75	2,23	3,55

$$MOS_{audio-predict} = -3,38 + (0,282 \cdot COL) + (0,467 \cdot NOI_a) + (0,7 \cdot DIC_a) + (0,345 \cdot LOU)$$

In Abbildung 4 kann man erkennen, dass die Qualitätsmodellierung, auch im audio-visuellen Kontext, sehr gut funktioniert. Die berechnet Pearson-Korrelation beträgt 0,97 und bestätigt den starken Zusammenhang zwischen den subjektiv ermittelten Sprachqualitätsbewertungen und den Werten die mittels des Modells geschätzt wurden. Auch der geringe „Root Mean Square Error“ (RMSE) von 0,26, zeigt die hohe Schätzgenauigkeit. Daraus lässt sich schließen, dass das abgegeben Urteil zur Sprachqualität auch über die zugrundeliegenden Sprachqualitätsdimensionen ermittelt werden kann. Weiterhin zeigt sich auch hier, dass das vorhandensein des Videokanals keinen Einfluss auf die Dimensionsbewertungen des Sprachkanals hat.

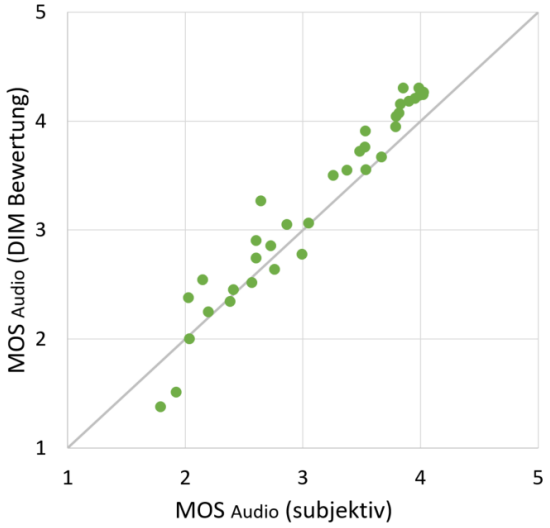

Abbildung 4: Ergebnis der dimesions-basierten Sprach-Qualitätsmodellierung.

Vergleich der Ankerkonditionen: Betrachtet man nun die Bewertungen der Anker-Testkonditionen, welche auch bereits im reinen Hörversuch untersucht wurden, zeigt sich auch hier, dass die Korrelation zwischen den Bewertungen sehr hoch ist (vgl. Tabelle 3 und Abbildung 5). Insbesondere bei der Dimension *Klangverfärbung* und *Rauschhaftigkeit* liegt sie bei über 0.9. Bei der Dimension *Diskontinuität* zeigt sich ein ähnliches Bild. Allerding fällt hier die Kondition $G.722.2 - PL0.02$ aus dem Rahmen, die im AV-Kontext deutlich besser bewertet wurde. Hier lässt sich annehmen, dass mitunter „vorteilhafte“ Stellen im Sprachkanal beeinträchtigt waren (z.B. Sprechpausen), so dass die Beeinträchtigung nicht zu hören ist. Aufgrund der Tatsache, dass nur 4 Proben im AV-Test mit dieser Störung behaftet waren, kann man hier keine definitive Aussage treffen. Für die Dimension *Lautheit* liegen für die Ankerkonditionen keine subjektiven Dimensionsbewertungen, aus einem reinen Hörversuch, vor. Dennoch weisen die Ergebnisse darauf hin, dass die Sprachqualitätsdimensionen auch im AV-Kontext stabil sind und sehr ähnliche Bewertugen erzielen.

Schlussfolgerung und Ausblick

Anhand des Experiment konnten wir zeigen, dass die Teilnehmer in der Lage waren die Sprachqualitätsdimensionen, auch bei verhandenem Video, zu bewerten. Weiterhin konnten wir zeigen, dass sich die Dimensionsbewertungen für die Sprachqualität nicht verändern, wenn der Videokanal gestört ist. Somit können die perzeptiven Qualitätsdimensionen aus der Sprachtelefonie auch im Videotelefonie-Kontext verwendet werden. Die Ergebnisse sind pausibel und ermöglichen das Schätzen der Gesamtqualität für den

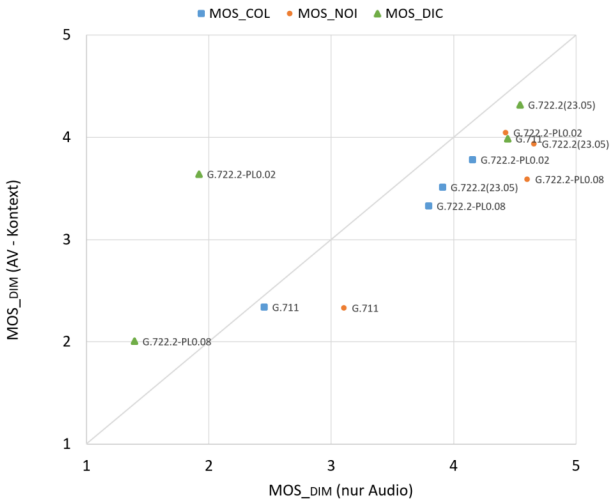

Abbildung 5: Vergleich der subjektiven Dimensions-MOS der Ankerkonditionen für die Dimensionen (COL - Klangverfärbung, NOI - Rauschhaftigkeit, DIC - Diskontinuität.

Tabelle 3: Vergleich der Dimensionsbewertungen der Ankerkonditionen für die Dimensionen (COL - Klangverfärbung, NOI - Rauschhaftigkeit, DIC - Diskontinuität), Pearson Korrelation.

Kondition	MOS_{COL}(AV)	MOS_{COL}(Audio)	Pearson
$G.711$	2, 34	2, 45	
$G.722.2(23.05)$	3, 51	3, 91	
$G.722.2 - PL0.02$	3, 78	4, 16	
$G.722.2 - PL0.08$	3, 33	3, 80	0, 99
Kondition	MOS_{NOI}(AV)	MOS_{NOI}(Audio)	Pearson
$G.711$	2, 33	3, 10	
$G.722.2(23.05)$	3, 94	4, 66	
$G.722.2 - PL0.02$	4, 05	4, 42	
$G.722.2 - PL0.08$	3, 59	4, 60	0, 94
Kondition	MOS_{DIC}(AV)	MOS_{DIC}(Audio)	Pearson
$G.711$	3, 99	4, 44	
$G.722.2(23.05)$	4, 32	4, 54	
$G.722.2 - PL0.02$	3, 64	1, 92	
$G.722.2 - PL0.08$	2, 01	1, 39	0, 83

Sprachkanal mit hoher Genauigkeit. Die Ergebnisse bestätigen weiterhin, dass die übliche Vorgehensweise bei der Schätzung der audio-visuellen Qualität, sinnvoll ist. Hierbei wird die Audio-/Sprachqualität und die Videoqualität getrennt von einander ermittelt und anschließend über eine Gewichtungsfunktion zum Gesamtqualitätsurteil zusammengefügt (vgl. bspw. [14]).
Um die Ergebnisse weiter zu festigen sind weitere Studien geplant.

Weiterhin wird an der instrumentellen Schätzung der Videoqualitätsdimensionen gearbeitet. Diese Schätzung wird anschließend mit der Schätzung der Sprachqualitätsdimensionen kombiniert. Aus diesem Ansatz soll ein wahrnehmungsbasierter Qualitätsschätzer entwickelt werden, um die audio-visuelle Gesamtqualität von Videotelefonie zu ermitteln.

Literatur

[1] ITU-T Rec. G.1070, *Opinion model for videotelephony applications*, International Telecommunication Union, CH-Geneva, 2012.

[2] M. Wältermann, *Dimension-based Quality Modeling of Transmitted Speech.* Springer-Verlag, D-Berlin, 2013.

[3] N. Côté, *Integral and Diagnostic Intrusive Prediction of Speech Quality*, ser. T-Labs Series in Telecommunication Services. Berlin, Heidelberg: Springer, 2011.

[4] F. Schiffner and S. Möller, "Direct scaling & quality prediction for perceptual video quality dimensions," in *2018 10th International Conference on Quality of Multimedia Experience (QoMEX).* IEEE - Signal Processing Society, 2018, pp. 51–53, electronic.

[5] ——, "Audio-Visuelle Qualität: Zum Einfluss des Audiokanals auf die Videoqualitäts- und Gesamtqualitätsbewertung," in *42. Jahrestag Für Akusitk (DAGA 2016).* Deutsche Gesellschaft für Akustik (DEGA e.V.), 2016, another medium.

[6] K. T. Inc., *Ishihara's Tests for Colour Deficiency - 24 Plates Edition*, Yushima Bunkyo-ku: Kanehara Trading Inc., 2015.

[7] H. Snellen, "Snellen chart," wikipedia – URL: https://commons.wikimedia.org/wiki/File:Snellen_chart.svg, letzter Besuch 08/2017.

[8] A. Raake, *Speech Quality of VoIP: Assessment and Prediction.* John Wiley and Sons, UK-Chichester, 2006.

[9] S. Möller, *Quality Engineering - Qualität kommunikationstechnischer Systeme.* Springer-Verlag, Heidelberg, 2010.

[10] ITU-T Rec. P.910, *Subjective video quality assessment methods for multimedia applications*, International Telecommunication Union, CH-Geneva, 04/2008.

[11] ITU-T Rec. G.191, *Software tools for speech and audio coding standardization*, International Telecommunication Union, CH-Geneva, 2010.

[12] F. Köster et al., *Comparison between the discrete ACR scale and an extended continuous scale for the qualtiy assessment of transmitted speech.* DAGA, D-Nürnberg, 2015.

[13] F. Friedemann, G. Mittag and S. Möller, "Modeling the overall quality of experience on the basis of underlying quality dimensions," in *2017 9th International Conference on Quality of Multimedia Experience (QoMEX 2017).* IEEE - Signal Processing Society, 2017.

[14] S. Möller and A. Raake, Eds., *Quality of Experience: Advanced Concepts, Applications and Methods*, 1st ed. Heidelberg, Germany: Springer, 2014, chapter 5.