Akustische Herausforderungen für interaktive Gruppenkommunikation

Janto Skowronek, Alexander Raake

TU Ilmenau – Institut für Medientechnik – Fachgruppe Audiovisuelle Technik, 98693 Ilmenau, Deutschland,

Email: janto.skowronek@tu-ilmenau.de

Einleitung

Moderne Sprachkommunikation geht heutzutage weit über das klassische Telefonieren hinaus. Insbesondere Mehrpersonenszenarien stellen eine stetig an Bedeutung gewinnende Art der Telekommunikation dar. Dabei stehen Entwickler vor neuen Herausforderungen um den steigenden Erwartungen der Nutzer gerecht zu werden.

Zu diesem Zweck verfolgt dieser Konferenzbeitrag das Ziel, neue Forschungsrichtungen in diesem Gebiet aufzuzeigen und anzustoßen. Dazu werden zunächst die anstehenden Herausforderungen aus zwei Perspektiven beleuchtet, nämlich hinsichtlich der Kommunikation und der Technik. Anschließend werden mögliche Lösungswege an einem Beispiel aus eigenen Arbeiten aufgezeigt.

Zwei Szenarien

Betrachtet man die akustischen Herausforderungen sowie die zwischenmenschlichen Kommunikationsprozesse, so muss zwischen zwei Arten von Mehrpersonenszenarien unterschieden werden, siehe Abbildung 1. In dem ersten Szenario, Szenario A, sind mehr als zwei Orte miteinander verbunden, und an jedem Ort befindet sich nur eine Person. In dem zweiten Szenario, Szenario B, sind zwei Orte miteinander verbunden, aber an mindestens einem Ort befinden sich mehr als eine Person.

Im Szenario A müssen die akustischen Probleme für je eine Sprecher- und Hörerposition im Raum gelöst werden, im Szenario B für mehrere Sprecher- und Hörerpositionen. Zudem repräsentiert das Szenario A ein Mehrpersonengespräch, welches ausschließlich über ein System läuft, während in Szenario B eine gemischte Kommunikation über ein System und Face-To-Face stattfindet.

Abbildung 1: Visualisierung der zwei betrachteten Szenarien von Mehrpersonengesprächen über ein Telekommunikationssystem.

Kommunikative Herausforderungen

In diesem Beitrag werden vier Aspekte der zwischenmenschlichen Kommunikation beleuchtet: Zweck, Inhalt, Prozess und Gesprächspartner.

Zweck Kommunikation verfolgt einen gewissen Zweck, welchen man grob in drei Kategorien einteilen kann. Diese wären das Erfüllen von Aufgabenstellungen, das Pflegen sozialer Kontakte und Bedürfnisse und der reine Informationsaustausch.

Das Erfüllen einer oder mehrerer definierter Aufgabenstellungen ist ein Zweck, der sich gut für Experimentalstudien eignet, da sich solche Aufgaben recht genau spezifizieren lassen, was die Reproduzierbarkeit bzw. die Vergleichbarkeit der einzelnen Kommunikationen fördert. Hierzu findet sich in der Literatur eine feinere Einteilung von Aufgabenstellungen nach McGrath[1], die häufig verwendet wird. McGrath unterscheidet demnach Aufgabenstellungen in vier Kategorien mit jeweils zwei Unterkategorien: *creativity* & *planning* (*generate*), *psychomotor* & *battles-contests* (*execute*), *mixed-motives* & *cognitive-conflict* (*negotiate*), *judgment* & *intellective* (*choose*).

Bezüglich der anderen beiden Kategorien ist insbesondere das Pflegen sozialer Kontakte interessant aufgrund der steigenden Anzahl an Gruppenkommunikation im privaten Umfeld, z.B. mit der entfernt lebenden Familie.

Inhalt Betrachtet man den Inhalt im Sinne des Sprachsignals, so erfordert eine erfolgreiche Kommunikation als Grundvoraussetzung eine ausreichende Sprachverständlichkeit. Die Literatur zum Thema Sprachverständlichkeit ist extrem umfangreich, als zwei Beispiele seien [2] und [3] genannt. Ein weiterer Aspekt hinsichtlich der Verbesserung von Systemen ist der Zusammenhang von Sprachverständlichkeit und Qualität, z.B. [4, 5].

Betrachtet man den Inhalt im Sinne der Bedeutung, so findet man in der Literatur den Begriff des Common Grounds. Er beschreibt die Tatsache, dass Gesprächspartner üblicherweise versuchen, ein gemeinsames Verstehen zu erreichen. Fussel [6] z.B. schreibt dazu, dass die Gesprächspartner ein gemeinsames Verständnis der Situation, der Aufgabe, sowie des jeweiligen Hintergundwissens, Erwartungen, Attitüden usw. suchen. Demnach sollten Systeme nicht nur reine Sprachverständlichkeit sicherstellen, sondern es den Gesprächspartnern auch ermöglichen, diese zusätzlichen Informationen aus dem Gespräch zu ziehen.

Prozess Betrachtet man den Ablauf einer Konversation, so ist ein wesentlicher Bestandteil das Turn-Taking, also das Wechselspiel zwischen den Sprechern. Ein prominentes Modell zum Turn-Taking wurde von Sacks et.al. [7] formuliert, welches davon ausgeht, dass Sprecherwechsel nur zu bestimmten Momenten während einer Konversation, den s.g. Transition-Relevance-Places, stattfinden. Demnach sollten diese nicht durch das System be-

einträchtigt werden, z.B. durch Signalstörungen zu genau diesen Zeitpunkten, oder Verschmierung dieser Zeitpunkte durch Verzögerung.

Ein weiterer Aspekt sind die s.g. Backchannels, nichtverbale Rückkopplungen der Zuhörer an den Sprecher, wie z.B. "hmm", "ja", oder "häh?". Diese Backchannels geben dem Sprecher ein Feedback, ob er verstanden wurde und weitermachen kann oder ob Bedarf zur Klärung oder eines Sprecherwechsels besteht. Demnach sollten Systeme diese oft recht leise geäußerten Feedbacksignale ebenfalls übertragen, welches bei einigen Verarbeitungsschritten wie Pegelwaagen oder Voice Activity Detection eine besondere Herausforderung sein könnte.

Und letztendlich sollte ein System auf höherer Ebene das Grounding unterstützen, mittels dessen die Gesprächspartner versuchen, einen Common Ground zu erreichen. Nach Clark & Brennan [8] besteht das Grounding aus einer Präsentationsphase einer Information (der Sprecher macht eine Äußerung) und einer Akzeptanzphase (der Hörer gibt Feedback darüber ob das Gesagte verstanden wurde). Dabei kann ein solcher Grounding-Rhythmus ein oder mehrere Turns umfassen, und das Feedback können Backchannels oder explizite Äußerungen, sprich eigene Turns sein.

Gesprächspartner Gesprächspartner machen sich während einer Konversation ein Bild der jeweils anderen Gesprächspartner. Dieses ist zunächst die Identität, welche bei Mehrpersonengesprächen über ein System ein wichtiger Faktor ist, und besonders im Zusammenhang von räumlicher Audiowiedergabe (Spatial Audio) untersucht wurde: die Fähigkeit, die Gesprächspartner mittels Spatial Audio auseinander zu halten, zeigte positive Effekte z.B. auf Focal Assurance Cues (Informationen wer die anderen Teilnehmer sind) [9], wahrgenommener Qualität und Konversationsaufwand [10], und kognitiver Anstrengung [11]. Neben der Fähigkeit, die Identitäten der Gesprächspartner auseinander zu halten, kann ein System aber auch die Wahrnehmung der Persönlichkeit insbesondere von unbekannten Gesprächspartnern beinflussen, welches z.B. durch Schoenenberg [12] für Signalverzögerungen gezeigt wurde. Und letztendlich tragen die Rolle und der Status aller Gesprächspartner zu einer Konversation bei, z.B. hinsichtlich einer formalen Gesprächsführung durch eine Führungsperson gegenüber einer offenen Diskussion gleichgestellter Teilnehmer.

Technische Herausforderungen

Bei der Betrachtung der technischen Herausforderungen liegt der Fokus in diesem Konferenzbeitrag auf die Signalqualität und möglichen Störeinflüsse. Die üblichen Einflüsse kann man grob anhand der Übertragungskette vom Sender zum Empfänger mit den Schritten Aufnahme, Übertragung und Wiedergabe betrachten, wobei eine derartige Zuordnung auch anderweitig gemacht werden kann. Wesentliche Aspekte der Aufnahme sind Pegel, Rauschen, Nachhall und Echo; Aspekte der Übertragung sind Verzögerung und Verzerrungen; und Aspekte der Wiedergabe sind Bandbreite und Lokalisation.

Aufnahme Für Szenario A müssen die Herausforderungen für eine Sprecherposition im Raum gelöst werden. Das heißt, hier kann man im wesentlichen auf die bereits existierenden Lösungen (Pegelwaage, Rausch-, Nachhall- und Echounterdrückung) zurückgreifen, insofern sich der Sprecher nicht in einer allzu schwierigen akustischen Umgebung befindet.

Für Szenario B müssen die Herausforderungen nun für mehrere Sprecher gelöst werden. Hier herrscht nach wie vor Forschungsbedarf, was auch die umfangreiche Literatur bezüglich der Signalverarbeitung (auch auf dieser Konferenz) zeigt. Offene Fragestellungen sind hier z.B. die Detektion multipler Sprecherpositionen, die Handhabe multipler Echopfade (insbesondere bei Mehrmikrofonlösungen), und vor allen Dingen unterschiedlich große Abstände zwischen Sprechern und Mikrofonen sowie den unterschiedlichen Einfallsrichtungen.

Übertragung Für Szenario A müssen die Herausforderungen bzgl. Verzögerung und Verzerrungen für multiple simultane Übertragungswege (je ein Weg pro Teilnehmerpaar) gelöst werden. Hier gibt es nachwievor offene Fragen, wie z.B. der Einfluss möglicher Verzerrungen durch die eingesetzte Mixtechnologie (Stichworte hier: Voice Activity Detection, Codec-Tandems, oder Select-And-Forward-Ansätze, siehe z.B. [13, 14, 15]), der Einfluss von Signalverzögerungen auf den Konversationverlauf (z.B. [16, 17, 12]), sowie der Aspekt der asymmetrischen Verbindungen, welcher unten am Beispiel aus eigenen Arbeiten genauer betrachtet wird.

Für Szenario B müssen die Herausforderungen für einen einzelnen Übertragungsweg gelöst werden. Während hier weitestgehend auf existierendes Wissen und vorhandene Lösungen zurückgegriffen werden kann, sind offene Fragen auch hier der Einfluss von Signalverzögerungen auf den Gesprächsablauf, oder der Einfluss von sich zeitlich verändernden Störungen (z.B. [18]).

Wiedergabe Für Szenario A müssen die Herausforderungen bzgl. der Wiedergabe für einen Zuhörer gelöst werden, während sie für Szenario B für mehrere Zuhörer an verschiedenen Positionen gelöst werden müssen. Als Herausforderungen können hier die Aspekte Bandbreite und Lokalisation genannt werden, wobei der Fokus in diesem Beitrag auf der Lokalisation einzelner Sprecher durch die Zuhörer liegt, welche durch neuartige Systeme mittels Spatial Audio ermöglicht wird.

Für Szenario A existiert bereits Wissen und Lösungen, welche vorwiegend mittels Kopfhörerwiedergabe realisiert sind; offene Fragen gibt es jedoch noch hinsichtlich einer möglichen Realisation durch Lautsprecherwiedergabe, der Frage nach einer möglichen Interaktion von Bandbreite und Spatial Audio [11], sowie der Interaktion von Spatial Audio und Übertragungsfehlern [19].

Für Szenario B besteht noch größerer Forschungsbedarf hinsichtlich der Lautsprecherwiedergabe. Zwar könnten hier z.B. Schallfeldsynthesesysteme zum Einsatz kommen; offene Fragen sind jedoch, welchen Effekt diese auf das Nutzererlebnis letztendlich haben und wie die Syste-

me entsprechend konfiguriert werden müssten.

Beispiel aus eigenen Arbeiten

Als Beispiel wird hier die Fragestellung der asymmetrischen Verbindungen für das Szenario A genauer betrachtet. Asymmetrische Verbindungen heißt, die einzelnen Teilnehmer haben unterschiedliche Verbindungs- oder Endgeräteeigenschaften, also z.B. ein Teilnehmer ist mit einem Mobiltelefon verbunden, die anderen mit Festnetzgeräten.

Die Fragestellung für so ein Szenario ist, ob man das Zusammenspiel dieser einzelnen Verbindungen hinsichtlich einer Gesamtqualität verbessern kann. Dieses führt zu einem Ansatz, den man auch unter Quality Engineering versteht, welcher aus drei Schritten besteht: 1. Untersuchen der Qualitätswahrnehmung, 2. daraus abgeleitet Entwicklung von Qualitätsmodellen, und 3. daraus abgeleitet die Systemsteuerung durch optimierte Einstellungen der technischen Parameter. Für das betrachtete Szenario hier bedeutet das, dass im ersten Schritt der Zusammenhang zwischen der Qualitätswahrnehumg der Einzelverbindungen (Individual Connection Quality Q_{ic}) und des gesamten Gruppenanrufs (Telemeeting Quality Q_i) untersucht werden muss; Details dazu werden in [20] veröffentlicht.

Im zweiten Schritt – der Modellentwicklung – wäre es sinnvoll, zunächst ein "perzeptives" Modell des Zusammenhangs $Q_i = f(Q_{ic})$ zu entwickeln, welches als Proof-of-Concept dient und im Idealfall sogar von bestimmten technischen Realisierungen unabhängig ist. Der Begriff "perzeptiv" bezieht sich hierbei auf den Aspekt, dass sowohl Eingangs- als auch Ausgangsgrößen des Modells Qualitätsbewertungen der Nutzer sind.

Als konkretes Beispiel sei hier ein Ergebnisplot aus [20] in Abbildung 2 gegeben. Die Abbildung zeigt die Schätzgüte verschiedener Modellierungsfunktionen in Form des Root-Mean-Square-Errors zwischen geschätzten und tatsächlich beobachteten Werten für Q_i. Grundlage der Ergebnisse ist ein Konversationstest mit 32 Drei-Personen-Gruppen, in dem verschiedene akustische und übertragungstechnische Störungen (Codec, Bandbreite, Rauschen, Echo, Paketverlust) dargeboten wurden. Während die vollständigen Details in [20] gegeben werden, sei hier der Fokus auf die Modellierungsfunktionen $F1$, $F4$, und $F7$ gelegt.

Funktion $F1$ ist der Mittelwert der Einzelverbindungen, d.h. $\hat{Q}_i = mean(Q_{ic})$; Funktion $F7$ der Minimumwert der Einzelverbindungen, d.h. $\hat{Q}_i = min(Q_{ic})$. Diese beiden Funktionen bilden die Modellierungsgrundlage, da sie nicht auf Daten trainiert werden müssen, und zudem äußerst intuitiv erscheinen: die Gesamtqualität ist einfach das Mittel der Einzelverbindungen, oder die Gesamtqualität ist durch die schlechteste Verbindung bestimmt.

Funktion $F4$ ist eine gewichtete Summe aus $F1$ und $F7$, wobei die Gewichte der beiden Terme zum Einem auf den Datensatz trainiert wurden, zum Anderen Sigmoid-Funktionen $1/(1 + exp(-x))$ sind, in denen wiederum der Mittelwert der Einzelverbindungen eingesetzt wurde. Dadurch wird die Gewichtung zwischen $F1$ und $F7$ qualitätsabhängig: je höher die Gesamtqualität, desto stärker wird sie durch den Mittelwert der Einzelverbindungen bestimmt, je niedriger die Gesamtqualität, desto stärker wird sie durch die schlechteste Einzelverbindung bestimmt. Damit ist diese Funktion sinnvoll interpretierbar, und sie ist laut Abbildung 2 außerdem auch noch tendenziell die beste Option: niedrigster Fehler, wenn auch die Unterschiede zu einigen anderen Funktionen nicht signifikant sind.

Zurückkehrend zu dem Quality Engineering, so wäre der nächste Schritt, diese Ergebnisse in ein instrumentelles Modell zu überführen. D.h. es müssen die Nutzerurteile für Q_{ic} durch instrumentelle Schätzungen $\hat{Q_{ic}}$ als Eingangsgrößen des Modells ausgetauscht werden, um ein solches Modell in der Praxis einsetzen zu können. Anschließend kann dann zum letzten Schritt übergegangen werden, in dem Optimierungsalgorithmen entworfen werden, die die technischen Parameter der Einzelverbindungen unter Zuhilfenahme des instrumentellen Qualitätsmodels so einstellen, dass eine bestmögliche Gesamtqualität erreicht wird.

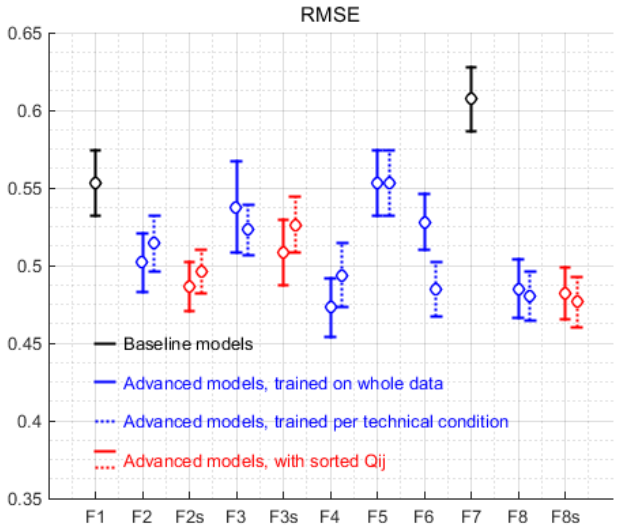

Abbildung 2: Resultate aus [20] für verschiedene Modellierungsfunktionen, aufgetragen als Root Mean Squared Error (RMSE) zwischen berechneten und tatsächlichen Qualitätsbeurteilungen. Details siehe Text.

Fazit

Abbildung 3 zeigt eine Übersicht, an welchen Schnittpunkten der zwei Perspektiven Kommunikation und Technik Forschung und Entwicklung stattfinden.

Aufnahme: Hier steht die Sprachverständlichkeit und damit der Inhalt im Vordergrund. Für Szenario A kann man einen Haken setzen, da hier viel Wissen und Lösungen existieren. Man sollte diesen jedoch in Klammern setzen, da nach wie vor Forschungsbedarf für akustisch schwierige Umgebungen herrscht. Für Szenario B sollte man wegen den wenigen Erkenntnissen dagegen ein Fragezeichen setzen.

Bandbreite (Wiedergabe): Hier kann man dieselbe Einschätzung wie oben geben.
Verzögerung (Übertragung): Hier gibt es einige Erkenntnisse hinsichtlich des Effektes auf den Prozess und die Wahrnehmung der Gesprächspartner. Zudem ist es bekannt, dass der Effekte von Verzögerung task-abhängig sind, sprich der Zweck wird hier ebenfalls betrachtet. Das verwendete Kreislaufsymbol deutet an, dass es viele Ergebnisse aber auch noch starken Forschungsbedarf gibt.
Verzerrung (Übertragung): Das angegebene Beispiel zeigte den Kenntnisstand sowie Forschungsbedarf für Szenario A auf, angedeutet durch das Kreislaufsymbol. Für Szenario B dagegen kann man weitestgehend auf vorhandenes Wissen und Lösungen zurückgreifen, was den in Klammern gesetzten Haken an dieser Stelle begründet.
Lokalisation (Wiedergabe): Die bestehenden Arbeiten hinsichtlich der Sprecheridentität und damit den Gesprächspartner zeigen, dass für Szenario A bereits Wissen aber auch noch Forschungsbedarf besteht, daher ein Kreislaufsymbol, während für Szenario B ein Fragezeichen angemessen ist.

Diese Einschätzungen sollen dazu anregen, die bestehenden Forschungsaktivitäten zu stärken bzw. neue zu starten. Aus Sicht der Autoren haben insbesondere drei Gebiete besondere Bedeutung: 1. die akustischen Herausforderungen hinsichtlich der Aufnahme insbesondere für Szenario B (mehrere Personen pro Seite), da eine ausreichenden Sprachverständlichkeit die Mindestvoraussetzung ist; 2. die übertragungstechnischen Herausforderungen für Szenario A (mehrere Endpunkte), da das Zusammenspiel der einzelnen Verbindungen und Endpunkte noch nicht vollständig geklärt und technisch umgesetzt ist, wie das angegebene Beispiel aufzeigte; 3. mittelfristig den Einfluss der technischen Systeme auf die Kommunikationsprozesse besser zu verstehen, um erfolgreiche Gruppenkommunikation über das Niveau einer ausreichenden Sprachverständlichkeit zu heben.

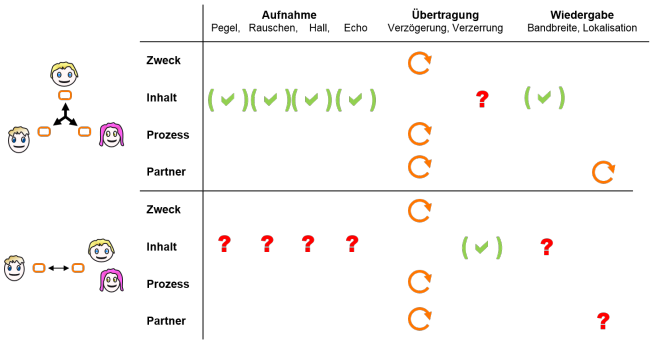

Abbildung 3: Überblick über die Forschungsaktivitäten anhand der zwei Perspektiven Kommunikation und Technik. Erläuterungen der Symbole siehe Text.

Literatur

[1] J.E. McGrath. Groups: Interaction and performance. Prentice-Hall Englewood Cliffs, NJ, 1984.

[2] C.H. Taal. Prediction and Optimization of Speech Intelligibility in Adverse Conditions. PhD thesis, Delft Technical University, 2013.

[3] S. Jørgensen. Modeling speech intelligibility based on the signal-to-noise envelope power ratio. PhD thesis, Technical University of Denmark, 2014.

[4] J. Preminger & D. van Tasell. Quantifying the Realtion on Between Speech Quality and Speech Intelligibility. In: Journal of Speech and Hearing Research 38 (1995), pp. 714–725.

[5] F. Schiffner, J. Skowronek, A. Raake, Speech Intelligibility and Quality of Transmitted Speech under Packet Loss, In: Proceedings of Forum Acusticum 2014.

[6] S.R. Fussell & N.I. Benimoff. Social and Cognitive Processes in Interpersonal Communication: Imlications for advanced telecommunicatinos technologies. In: Human Factors 37 (1995), pp. 228–250.

[7] H. Sacks, E.A. Schegloff, & G. Jefferson. A Simplest Systematics for the Organisation of Turn-Taking for Conversation. In: Language 50 (4 Dec. 1974), pp. 696–735.

[8] H.H. Clark and S.E. Brennan. Grounding in Communication. In: Perspectives on socially shared cognition. American Psychological Association, 1991, pp. 127–149.

[9] J.J. Baldis. Effects of Spatial Audio on Memory, Comprehension, and Preference during Desktop Conferences. In: Proceedings of the ACM CHI 2001 Human Factors in Computing Systems Conference, 2001, pp. 166–173.

[10] A. Raake, et al. Listening and conversational quality of spatial audio conferencing. In: Proceedings of the AES 40TH International Conference. 2010.

[11] J. Skowronek & A. Raake. Assessment of Cognitive Load, Speech Communication Quality and Quality of Experience for spatial and non-spatial audio conferencing calls. In: Speech Communication (2015), pp. 154–175.

[12] K. Schoenenberg. The Quality of Mediated-Conversations under Transmission Delay. PhD Thesis. Technische Universität Berlin, 2015.

[13] K. Singh, G. Nair, & H. Schulzrinne. Centralized conferencing using SIP. In: Internet Telephony Workshop. 2001, pp. 57–63.

[14] P.J. Smith, P. Kabal, & R. Rabipour. Speaker Selection for Tandem-Free Operation VoIP Conference Bridges. In: Proceedings of IEEE Workshop on Speech Coding. 2002, pp. 120–122

[15] S. Firestone, T. Ramalingam, & S. Fry. Voice and Video Conferencing Fundamentals. Cisco Press, 2007.

[16] P.T. Brady. A statistical analysis of on-off patterns in 16 conversations. In: Bell Syst. Tech. J. 47.1 (1968), pp. 73–99.

[17] F. Hammer. Quality aspects of packet-based interactive speech communication. PhD thesis, Technical University Graz, 2006.

[18] Benjamin Weiss et al. Temporal Development of Quality of Experience. In: Quality of Experience - Advanced Concepts, Applications, Methods. Springer, 2014, pp. 133–147.

[19] M. Spur, D. Guse, J. Skowronek. Influence of Packet Loss and Double-Talk on the Perceived Quality of Multi-Party Telephone Conferencing with Binaurally Presented Spatial Audio Reproduction. Tagungsband der DAGA 2016.

[20] J. Skowronek, Quality of Experience of Multiparty Conferencing and Telemeeting Systems - Methods and Models for Assessment and Prediction. PhD thesis draft, to be submitted in 2016, Technische Universität Berlin.