IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 12, 2022, accepted May 5, 2022, date of publication May 18, 2022, date of current version June 21, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3176369

Quality of Experience in Telemeetings and Videoconferencing: A Comprehensive Survey

JANTO SKOWRONEK[1], ALEXANDER RAAKE[2], (Member, IEEE), GUNILLA H. BERNDTSSON[3], OLLI S. RUMMUKAINEN[4], (Member, IEEE), PAOLINO USAI[5], SIMON N. B. GUNKEL[6], (Senior Member, IEEE), MATHIAS JOHANSON[7], EMANUËL A. P. HABETS[4], (Senior Member, IEEE), LUDOVIC MALFAIT[8], DAVID LINDERO[9], AND ALEXANDER TOET[10], (Life Senior Member, IEEE)

[1]Hochschule für Technik Stuttgart–University of Applied Sciences, 70174 Stuttgart, Germany
[2]Audiovisual Technology Group, Ilmenau University of Technology, 98693 Ilmenau, Germany
[3]Ericsson Research, 164 80 Kista, Sweden
[4]International Audio Laboratories Erlangen, 91058 Erlangen, Germany
[5]European Telecommunications Standards Institute (ETSI), 06921 Sophia Antipolis, France *(Retired)*
[6]TNO—Netherlands Organization for Applied Scientific Research, 2509 JE The Hague, The Netherlands
[7]Alkit Communications AB, 431 37 Mölndal, Sweden
[8]Dolby Laboratories Inc., Sunnyvale, CA 94085, USA
[9]Ericsson Research, 977 53 Luleå, Sweden
[10]TNO—Netherlands Organization for Applied Scientific Research, 3769 ZG Soesterberg, The Netherlands

Corresponding author: Janto Skowronek (janto.skowronek@hft-stuttgart.de)

The work of Janto Skowronek was supported by the M4_LAB-HFT-Innovationslabor für die Metropolregion 4.0 (M4_LAB) Project by the German Federal Ministry of Education and Research under Grant 03IHS032A. The work of Alexander Raake was supported by the Cyberphysische Telepräsenzsysteme zur Erforschung der Mensch-zu-X-Kommunikation (CYTEMEX) Project by the Free State of Thuringia, Germany, under Grant FKZ: 2018-FGI-0019.

ABSTRACT Telemeetings such as audiovisual conferences or virtual meetings play an increasingly important role in our professional and private lives. For that reason, system developers and service providers will strive for an optimal experience for the user, while at the same time optimizing technical and financial resources. This leads to the discipline of Quality of Experience (QoE), an active field originating from the telecommunication and multimedia engineering domains, that strives for understanding, measuring, and designing the quality experience with multimedia technology. This paper provides the reader with an entry point to the large and still growing field of QoE of telemeetings, by taking a holistic perspective, considering both technical and non-technical aspects, and by focusing on current and near-future services. Addressing both researchers and practitioners, the paper first provides a comprehensive survey of factors and processes that contribute to the QoE of telemeetings, followed by an overview of relevant state-of-the-art methods for QoE assessment. To embed this knowledge into recent technology developments, the paper continues with an overview of current trends, focusing on the field of eXtended Reality (XR) applications for communication purposes. Given the complexity of telemeeting QoE and the current trends, new challenges for a QoE assessment of telemeetings are identified. To overcome these challenges, the paper presents a novel Profile Template for characterizing telemeetings from the holistic perspective endorsed in this paper.

INDEX TERMS Audio, extended reality, quality, quality of experience, teleconferencing, telemeetings, video, videoconferencing.

I. INTRODUCTION

More than 150 years after the invention of the telephone, state-of-the art features such as video transmission and screen sharing prove that today's telecommunication technology has evolved well beyond mere speech-based, audio-only communication. With the advent of modern eXtended Reality (XR) technologies (i.e., virtual, mixed, or augmented reality) even more natural or more immersive telecommunication experiences are possible.

Human-to-human interaction over a telecommunication system, also referred to as mediated communication, is part of our daily life, both in professional and private contexts. Considering the different societal, economic, climate and

The associate editor coordinating the review of this manuscript and approving it for publication was Mauro Fadda.

technological changes over the last couple of decades, the relevance of such systems is still increasing.

Moreover, users are confronted with many different technical possibilities to communicate remotely, but are often experiencing high cognitive load and fatigue during such mediated communication sessions (see e.g., [1]). Accordingly, the demand for high-quality mediated communication is large and increasing, which in turn translates into system quality requirements, both from a service provider's and user's perspective.

In this context, a systematic analysis of the Quality of Experience (QoE) [2] of telecommunication systems, as it is perceived by the user, can help developers and service providers to improve their solutions and services. The notion of quality was already included in the old patents on telephone technology. For example, according to Richards [3], the original patent by Edison from 1877 stated that the carbon microphone was much better sounding than the initial design by Bell from 1876. Since then, quality and ultimately QoE assessment have developed to a well established discipline in the telecommunications sector (see e.g., [4]).

However, with the developments mentioned above, new systems bring additional challenges and opportunities for both the user and the service provider. Therefore, existing QoE assessment approaches need to be continuously extended to new types of systems as well as new user expectations. For that reason, academic and industrial research as well as telecommunication standardization bodies are highly active, not only in developing new telecommunication solutions but also in developing corresponding new QoE assessment methodologies.

A. CONTRIBUTIONS OF THE PAPER

This paper provides the reader with an entry point to this comprehensive and still growing field of QoE assessment of mediated communication, with a focus on modern and near-future telemeeting systems as defined in Section II-B. To this aim, a structured survey of relevant scientific literature is presented, systematically considering a large number of aspects that are needed to understand the QoE of telemeetings.

To illustrate the different aspects of telemeetings, Figure 1 visualizes a telemeeting system with its technical main components, connecting multiple participants with different intentions, emotions and expertise, who are situated in different environments with different physical objects relevant for the meeting.

The paper makes a significant contribution to the state of the art by means of five individual contributions. The first three contributions are associated with the survey approach taken, which considers three different angles, each of which is explored in detail in the paper:

(1) Analysis and structuring of telemeeting assessment literature in terms of the associated Quality Influence Factors (QIFs), reflecting the approach now widely adopted in the QoE community initially suggested in [2], [5], see Section IV.

TABLE 1. Table of contents.

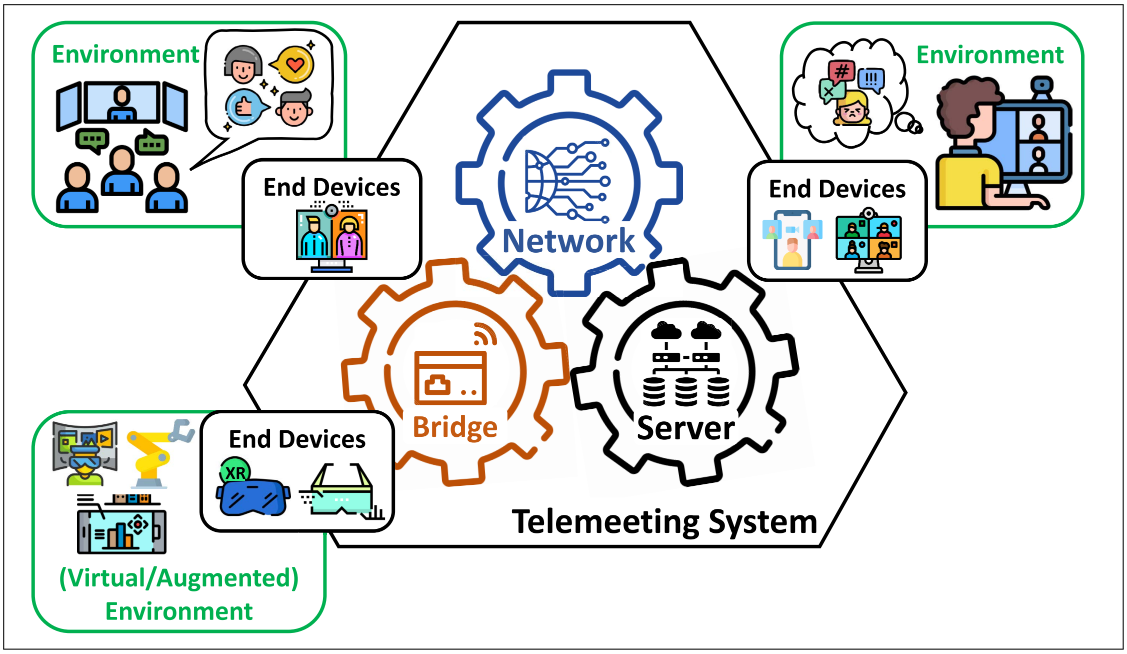

FIGURE 1. Visualization of an exemplary telemeeting system with its main components network, conferencing bridge, server infrastructure, and end devices. The system connects three sites, each characterized by a certain number of participants and the respective physical or virtual/augmented environment. The icons visualize the various technical possibilities, different meeting contexts, and human aspects that play a role in telemeetings.

(2) Analysis and structuring in terms of communication and perception processes in light of QoE, integrating the views on the QoE formation process [2], [6]–[8] and on human communication and conversation analysis [9], see Section V. Here, the paper takes a rather novel approach by putting an emphasis on inter- and intrapersonal communication processes and how they relate to QoE (Section V-A); followed by a synopsis of perceptual and cognitive processes that are relevant for QoE (Section V-B).

(3) Analysis and systematization of evaluation methods for telemeetings, including audio, video and interaction assessment, and considering both perceptual, "subjective" and instrumental, "objective" methods, see Section VI.

Moreover, there are two further contributions of this paper:

(4) Embedding QoE assessment of telemeetings into recent technological trends and expected new types of services by means of a survey of new, XR-based telemeeting technology and its implications for QoE, see Section VII.

(5) Introduction of the *Telemeeting Profile Template*, that is, a tool that provides a set of quantifiable criteria for telemeeting QoE evaluation. This allows to systematically and holistically characterize telemeetings from a QoE perspective, and select the right approach for a given assessment task, see Section VIII.

Moreover, the paper takes the complexity of QoE (see Section II-A) into account, by deliberately discussing both technical and non-technical aspects. This holistic perspective taken by the paper addresses both researchers and practitioners. Based on their multidisciplinary expertise, the authors are convinced that this holistic presentation will help the technical experts to further improve telemeeting systems or to develop new methods, in particular for a technology-oriented assessment of telemeeting QoE, while well reflecting the different application-scenarios and hence user and context factors. Along this line of thought, the paper identifies relevant, possible links between non-technical aspects and potential or proven technical approaches, to consider these for further system or evaluation-method development.

Summarizing, it is argued that approaching telemeeting QoE from a holistic perspective has a number of benefits that foster further progress in both technology and QoE assessment of telemeetings.

B. STRUCTURE OF THE PAPER

Table 1 shows the structure of the rest of this paper, which consists of eight main sections and a number of respective subsections, that are organized as follows. In Section II, the paper starts with a background on terms and concepts of telemeeting QoE. Then, Section III provides a detailed explanation of the process that the authors used to conduct the survey. Next, each of the following sections form one of the five contributions of the paper. First, Section IV structures relevant mediated communication quality aspects in terms of QIFs. Second, Section V takes a closer look at the QoE-relevant processes. Third, Section VI provides an overview of current methods for telemeeting QoE evaluation. Fourth, Section VII gives an overview of current development trends in telemeetings, focusing on XR-based technology.

Fifth, Section VIII presents a new approach for structuring telemeeting QoE assessment. For that purpose, a Telemeeting Profile Template is proposed to streamline the knowledge from the previous survey sections, indicating how the large body of QoE-related aspects can be applied for a holistic QoE evaluation of current and future telemeetings. Finally, a few closing remarks in Section IX conclude the paper.

II. TERMS AND CONCEPTS

A. DEFINITION OF QUALITY OF EXPERIENCE (QoE)

Based on [2], Quality of Experience (QoE) has been defined by the ITU-T as *"The degree of delight or annoyance of the user of an application or service"* [10]. That means, QoE is a construct that is formed inside a person's mind, see e.g., [2], [6], [8], and is based on the person's experience with an application, service or event, which here is a telemeeting. Accordingly, QoE is to be considered as a complex cognitive construct, resulting from technical aspects of a telemeeting system, but strongly influenced also by numerous other human and contextual aspects, see e.g., [2], [5]. The paper reflects this complexity by providing a systematic approach towards a holistic perspective on telemeeting QoE as already described in Section I-A.

B. DEFINITION OF TELEMEETINGS

Contemporary telemeeting systems can be realized in numerous ways, ranging from telephone conference bridges over audiovisual computer-based solutions and high-end telepresence rooms to systems using virtual, mixed, or augmented reality (often referred to as eXtended Reality, XR). The solutions can differ in various aspects such as specific devices, transmission technologies, collaboration and management features, and more. Correspondingly, a variety of terms are used to refer to such systems.

While those terms often give a principal idea about the system characteristics in general, they are hardly formally defined and sometimes even inconsistently used across contexts. For example, the difference between a telephone and a telepresence room is rather clear – at least for people that have used both systems. In turn, the term *conferencing system* usually refers to multiparty communication scenarios, while the term *video conferencing system* is also often used for a one-to-one video communication system.

To account for such aspects and to have a single term summarizing such different telecommunication systems, the term *telemeeting* is promoted by the International Telecommunication Union (ITU). A dedicated work group of ITU-T, ITU's telecommunciation sector, addresses telemeeting QoE, namely *Study Group 12 – Question 10*. In its task description [11], telemeeting is used

> *to cover with one term all means of audio or audiovisual communication between distant locations.*

This is similar to the formal definition given in [12]:

> *A meeting or conference at which people in different locations participate by means of telecommunications technology.*

While this term is quite encompassing for all kinds of telecommunication systems, the two mentioned definitions suggest some limitations. These limitations are clarified in the following to better specify the scope of this paper.

First, a telemeeting uses speech as the primary communication modality, which then may be augmented by other modalities such as video, images, text, or in case of virtual or augmented reality also haptic and olfactory information. Thus, any sole means of communication without speech are not considered as a telemeeting system. An exception to this are telemeeting systems for hearing impaired people, which use text or video to replace the missing audio channel.

Second, to qualify as a telemeeting, the system should allow for bidirectional communication between the participants, meaning that unidirectional transmission as in radio or television broadcasts are not considered as telemeetings.

Third, a telemeeting is a real-time communication session between participants, sometimes also referred to as synchronous communication. In contrast, means of asynchronous communication, such as email, sending a text, audio or audiovisual messages, or social media posts, are not considered as telemeeting here. As every telecommunication system has a certain end-to-end transmission delay between participants, the qualifier *real-time* need not be interpreted as instantaneous or zero-delay. This means that systems with a certain delay are still considered as telemeeting systems here, as long as the system's original purpose is to approximate synchronous communication, even if the delay is noticeable by the participants or may even trigger different communication behavior compared to a truly instantaneous communication.

These three qualifiers – speech as the primary communication modality and bidirectional, real-time communication – define the scope of telemeeting systems in this paper. Accordingly, the following definition is proposed (adapted from [12]):

> *A telemeeting is a meeting or conference at which people in different locations participate by means of telecommunication technology, and exchange speech-based audio or audiovisual information, in a bidirectional and synchronous, real-time manner.*

This definition highlights the primary function of telemeetings, which is the exchange of information between participants. However, the definition should not be understood as a limitation of the purpose of a telemeeting. In fact, telemeetings can serve a variety of purposes ranging from the exchange of information up to the real-time collaboration on a certain set of tasks or merely social interaction. For more details on purposes, see also Section IV-B3. Such a broader perspective on telemeetings beyond information exchange is taken in this paper for two reasons. First, this perspective accounts for the technical possibilities of modern systems and prototypes. Examples are state-of-the-art collaboration features such as virtual whiteboards and joint document editing, as well as the potential of XR technologies, remote sensing and control of physical objects as in cyber-physical systems. Second, this perspective accounts for modern working styles, in which groups directly collaborate during a meeting and produce results right away, which is reflected in the discipline of remote and computer-supported collaborative working.

C. TELEMEETINGS AND FACE-TO-FACE MEETINGS

One of the most fundamental questions regarding the QoE of telemeetings is: what is the relation between mediated communication during telemeetings and face-to-face meetings? A conventional approach is to directly compare mediated communication and face-to-face meetings, often setting the face-to-face meeting as the ultimate goal or reference. This approach is suitable in many contexts and has been included in a highly formalized method for measuring task performance in telemeetings [13], which was developed with contributions from the authors.

However, over the last decades research has shown that a face-to-face meeting is not always the best form of group communication. For instance, see the extensive literature analyses considering more than 200 peer-reviewed conference papers and journal articles in [14]–[17] or theoretical work such as the Task Media Fit Model [18].

An alternative, more open perspective found in the literature, and also taken in this paper, is to acknowledge mediated communication in telemeetings as its own way of communication, and to keep a differentiated view at it throughout. Then, a fair comparison between telemeetings and face-to-face meetings requires a good understanding of the exact circumstances and precise specification of the concrete aspects considered.

On the one hand, such a perspective allows to better understand processes and aspects that contribute to a good QoE of telemeetings. On the other hand, such a perspective enables to include future solutions in which face-to-face meetings are not just mimicked but new communication and collaboration experiences are created.

III. METHOD

One underlying goal for conducting this survey was to create a scientific basis for the Telemeeting Profile Template, a tool for a systematic characterization of telemeetings from a QoE perspective, see also Section VIII. In that respect, the literature survey was conducted hand in hand with the development of said tool as follows:

The author team compiled relevant literature in an iterative process, combining a bottom-up and top-down approach. The goal of this procedure was to obtain a comprehensive list of relevant aspects from the literature and practical experience (bottom-up path), and to identify a structure for this list of aspects (top-down path).

As part of the bottom-up path, the authors compiled a set of individual aspects that are relevant from a QoE perspective. In the paper, these are referred to as Quality Influence Factors (QIFs), cf. e.g., [2], [5]. The authors used different sources for compiling respective literature: First, all members of the author team included citations used in earlier work in the field that each co-author was aware of, based on their individual long-term expertise in the field. Second, dedicated literature search queries were conducted for each factor that was not already covered by the first compilation of literature, or when the authors saw the need to better understand a factor. Last, whenever feasible, the authors traced back original papers that were cited in above mentioned sources.

For some factors the scientific evidence was quite strong and direct, e.g., when studies found that a factor influences quality ratings. For other factors the evidence was more indirect in the sense that a factor has been shown to influence the communication, which in turn influences QoE. Accordingly, publications showing such a direct or indirect relevance of a factor on QoE were included in this survey; corresponding references can be found in Tables 2 to 5 in the Column *Quality Relevance*. For further factors, the scientific evidence was less clear, e.g., when the aspect or similar terms have been mentioned in the reviewed literature, but little more information or evidence with respect to QoE was given. In many such cases, the author team identified background information which could serve as a starting point for further study; corresponding references can be found in Tables 2 to 5 in the Column *Background*. However, there were still many factors that the authors considered as highly relevant from a practical experience, but for which no dedicated scientific literature was found; those cases are indicated by the "—" symbol in Tables 2 to 5.

As part of the top-down path, the authors had regular telemeetings in which they discussed a possible structure and the completeness of the list of factors, starting from the three main categories of QIFs and building on the different expertise of each co-author, to account for the fact that telemeetings can differ in various aspects. In addition, the authors collected feedback on intermediate versions from further experts from ITU-T Study Group 12 [19], a standardization group working on Quality of Experience and Quality of Service in the telecommunication sector. This feedback served to refine the list as well as to confirm the practical relevance of those aspects for which little scientific evidence was found.

As explained above, much of the literature search for this survey was conducted with a focus on QIFs (Section IV). The combination of domain knowledge of the author team and dedicated literature search on individual aspects was also used for the remainder of the paper, i.e., the survey on QoE-relevant processes (Section V), the overview of widely adopted assessment methods of telemeetings (Section VI), and the discussion of current trends concerning telemeetings (Section VII).

IV. SURVEY ON QUALITY INFLUENCE FACTORS (QIFs)

To structure the numerous aspects QoE is influenced by, the concept of *Quality Influence Factors (QIFs)* with the three main categories *Human Influence Factors (HIFs)*, *System Influence Factors (SIFs)* and *Context Influence Factors (CIF)* has been introduced. In [2], a QIF is defined

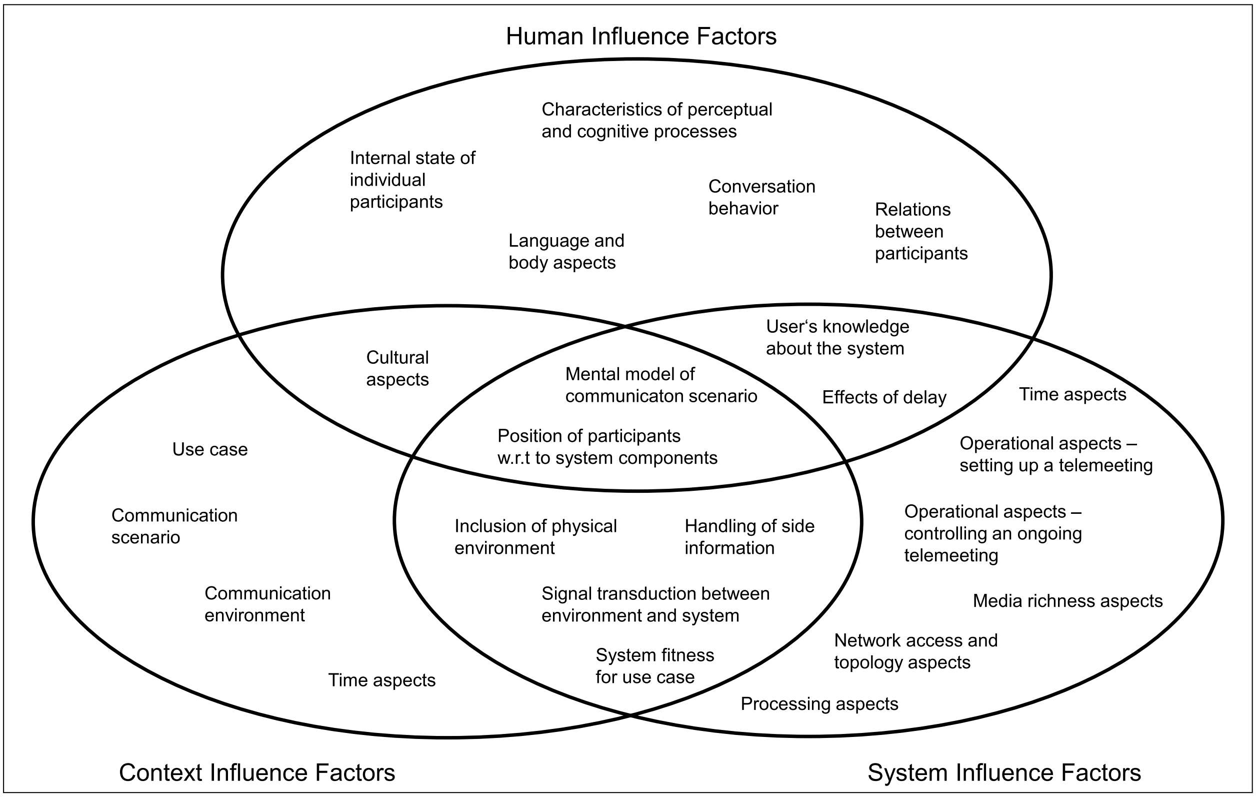
Human Influence Factors
Characteristics of perceptual and cognitive processes
Internal state of individual participants
Conversation behavior
Relations between participants
Language and body aspects
Cultural aspects
User's knowledge about the system
Mental model of communicaton scenario
Effects of delay
Time aspects
Position of participants w.r.t to system components
Operational aspects – setting up a telemeeting
Use case
Communication scenario
Inclusion of physical environment
Handling of side information
Operational aspects – controlling an ongoing telemeeting
Media richness aspects
Communication environment
Signal transduction between environment and system
System fitness for use case
Network access and topology aspects
Time aspects
Processing aspects
Context Influence Factors
System Influence Factors

FIGURE 2. Visualization of the different categories and sub-categories used to structure the telemeeting-relevant Quality Influence Factors (QIFs). The categories are expressed as the three different types of QIFs (Human, Context and System Influence Factors) as well as their combinations. Each category can contain a number of sub-categories to further structure the list of QIFs. See text for more details.

as: *"Any characteristic of a user, system, service, application, or context whose actual state or setting may have influence on the Quality of Experience for the user."* Reiter *et al.* [5], for instance, discuss a number of factors for QoE assessment in general; Akhtar and Falk [20] briefly summarize QIFs that should be considered in audiovisual multimedia quality assessment; Bouraqia *et al.* [21] give an overview of QIF for video streaming applications, and Seufert *et al.* [22] provide a more detailed taxonomy of QIFs for HTTP-based adaptive streaming technology. Highly relevant for telemeetings and this paper are the publications by Husić *et al.* [23], who give an overview of QIFs for unified communication systems (i.e., integrated services combining telemeeting functionality with asynchronous communication means), and Vučić and Skorin-Kapov [24], who review a number of QIFs in the context of mobile audiovisual telemeetings. As all these publications show, the list of possible QIFs can become quite long. For that reason, this paper provides a more detailed structure based on the three main categories of QIFs as follows.

To start with, it should be noted that the three main categories of QIFs are not always fully separable, see for instance [5]. For that reason, the proposed structure also allows for categories that represent two or even all three types of influence factors, which are here referred to as *Mixed Influence Factors (MIF)*.

To further structure the list, a second grouping hierarchy was added using sub-categories. Individual QIFs are grouped into these sub-categories whenever they share some aspects, such as the *Internal state of individual participants* as a sub-category of the group *Human Influence Factors*.

Figure 2 represents the categories and the sub-categories in a Venn-type diagram that visualizes the possible overlap of the QIFs. Note that not all individual factors are plotted here, to enable better readability. Instead, the full list of QIFs is given in Tables 2 to 5, while even more detailed information about the individual factors can be found in the supplementary material of this paper [25], [26]. The next four sections provide a survey of respectively System, Context, Human and Mixed QIFs.

A. SYSTEM INFLUENCE FACTORS (SIFs)

System Influence Factors (SIFs) refer to the technical characteristics of a system that influence QoE. According to the Qualinet white Paper [2], SIFs can relate to content, media, network and devices, and refer, for example, to aspects ranging from signal capture over transmission to reproduction. This kind of signal processing perspective will be taken in Section IV-A1. Additional SIFs, which refer to technical characteristics concerning the user interaction with the system, are addressed in Section IV-A2.

TABLE 2. Overview of System Influence Factors, organized along the sub-categories of Figure 2. See text for details.

Category: System Influence Factors

Sub-Category	Quality Influence Factor	Quality Relevance	Background
Media richness aspects	Communication mode	[14]–[17], [27] [28, Chap. 2]	[29]–[31]
	Auditory representation of participants	—	—
	Auditory representation of environment	—	—
	Visual representation of participants	—	—
	Visual representation of environment	—	—
	Spatial audio reproduction	[32]–[35]	[36]
	Spatial video reproduction	[37], [38]	—
	Virtual Shared Workspace	[39], [40]	—
	Degree of Realism	[41], [42]	—
	Degrees of Freedom	—	—
Processing aspects	Audio mixing paradigm	—	[43]–[45]
	Video mixing paradigm	—	[46], [47]
	Audio signal processing - coding technology	[7], [48]–[52]	[36], [52]–[57]
	Audio signal processing - signal enhancement	[48]–[50]	[52], [54], [58]–[60]
	Video signal processing - coding technology	[48]–[50], [61], [62]	[63]–[68]
	Video signal processing - signal enhancement	[48]–[50]	[63], [69]–[72]
Time aspects	Audiovisual asynchrony	[73]–[75]	[76], [77]
	End-to-end delay	[73], [74], [78]–[83]	—
Network access and topology aspects	Behavior of network access	[28]	[84]
	Computation distribution (determined by topology)	—	[46], [47]
Operational aspects - setting up a telemeeting	Call setup	—	—
	Participant registration	[85]	—
	Typical problems when setting up a telemeeting	[86], [87]	—
	Installation complexity - number and type of components	[88]–[90]	—
	Installation complexity - configuration	[89], [90]	—
	Installation complexity - signal calibration	—	—
Operational aspects - controlling an ongoing telemeeting	User interface presentation modality	[91]–[93]	[94]
	User interface control modality	[89], [91], [92]	—

Notes: The column *Quality Relevance* cites either empirical studies directly investigating the factor's effects or publications discussing the relevance more from a theoretical point of view. Moreover, the relevance can refer either directly to QoE or to perception or communication aspects, which in turn are relevant for QoE. When no references are given, the factor is considered to be relevant by the authors, yet requires further study for scientific proof. The column *Background* provides pointers to further literature.

1) SIFs RELATED TO THE SIGNAL TRANSMISSION OVER THE SYSTEM

This section outlines the general processing and transmission stages for the audiovisual signals between the different connected sites of a telemeeting. In that respect, this section refers to the following sub-categories of SIFs according to Table 2: Media richness aspects, Processing aspects, Network access and topology aspects, Time aspects. Since many different instantiations of telemeeting systems are possible, a generalized perspective is taken here. As the paper addresses quality and QoE, the descriptions in this section focus on the question of *"What is happening to the information along the way between participants?"* rather than on the question of *"How are the processing and transmission steps realized?"*

A typical approach in communication and media technology is to consider the end-to-end chain as a channel from the source/sender to the sink/receiver. In the case of interactive

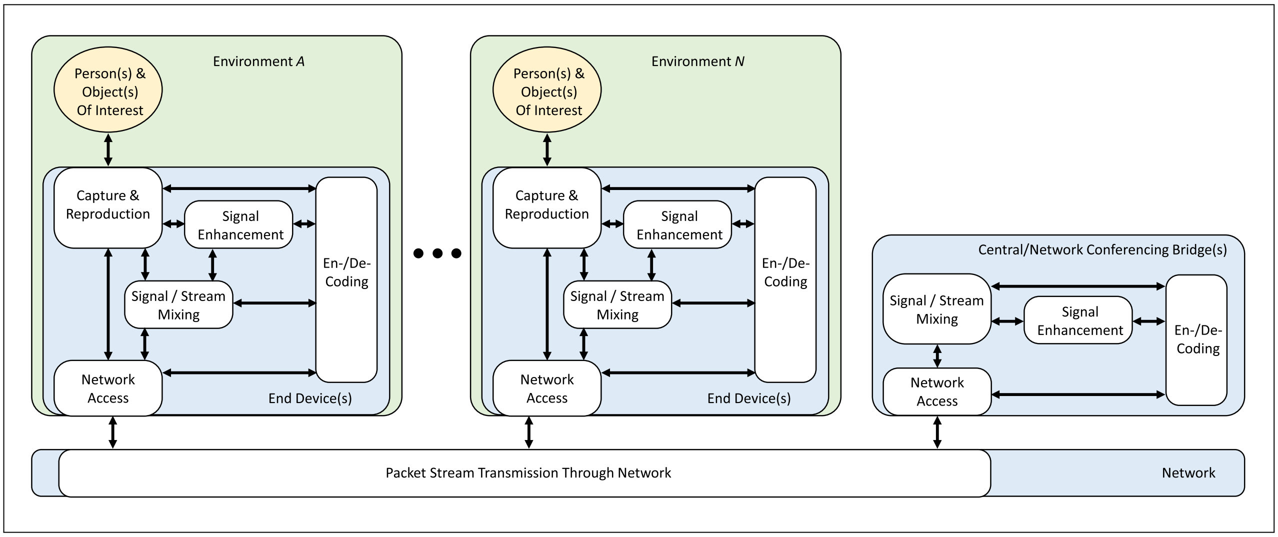

FIGURE 3. Generalized overview of end-to-end transmission chain between *N* sites, the network and any involved network conferencing bridges. For details see Section IV-A1.

two-party scenarios like in traditional telephony, this representation typically includes the end-to-end chain in both directions and any signal paths between them. This allows to include interaction-related aspects such as the impact of transmission delay or signal processing stages that require the consideration of signals in both send and receive directions such as echo cancellation. Examples for telephony are given in [7], [95] and [96], for video telephony in [97]. In [98] and [28, Chap. 6.3], the authors also extended such considerations for multiparty scenarios and proposed an approach to analyze such multiparty settings in more detail and from a QoE perspective.

In this paper, a simplified view is used: Figure 3 visualizes the major components of a telemeeting system connecting *N* sites. At each site, one or more persons are located in a certain environment. Moreover, a number of additional objects of interest may be present in one or more of these environments, such as a physical whiteboard or a physical object that is the topic of the discussions, such as a prototype system or the like. From a QoE perspective, a relevant factor is the degree to which wanted as well as unwanted information from the participants, about the objects of interest and about the environment is transmitted over the system. This leads to a number of SIFs that are related to the media richness provided by the system, such as the auditory and visual representation of participants and environments.

To connect each site to the telemeeting, one or more end devices may be used. The end devices perform a number of processing steps, which are subsumed as a SIF in the sub-category *Processing aspects*. In modern telemeeting solutions, these steps usually consist of a number of subprocesses, which in addition are often interlinked. For example, the encoding and decoding of signals often combines signal processing with networking-specific mechanisms, and it can be carried out at different places, e.g., as part of the capture & reproduction, the signal enhancement, the network access, or the mixing stages.

Moreover, not all components are used in all system instances. For example, conventional telephone conference bridges use a central conferencing bridge on a server in the network, that is, no mixing blocks are needed in the end devices. In another example, the system connects the *N* sites using a client-side bridging technology, omitting the need for a central mixing bridge. In this case, the signal enhancement steps might be connected not only to the capture, reproduction and coding steps, but to mixing steps as well.

As these examples illustrate, differences between individual systems can be quite large, which makes it hard to come up with a general picture at a more detailed level. For that reason, Figure 3 simplifies this by showing five major processing steps in the end device: *Capture & Reproduction*, *Signal Enhancement*, *En-/De-Coding*, *Signal / Stream Mixing* (in case of a peer-to-peer system as it requires client-side conferencing bridges), and *Network Access*. There is a vast amount of literature on the available technologies, ranging from electro-acoustic and electro-optic transducers (e.g., [69], [99]) over signal processing algorithms for signal enhancement and data compression (e.g., [36], [53]–[60], [63]–[68], [70]–[72]) to data error correction mechanisms for the packet streams (e.g., [100], [101]). Next to research work focusing on the technology, a number of publications discuss these technical aspects of telemeeting and telecommunication systems from a QoE perspective, e.g., [7], [51], [52], [61], [62], [95]. Moreover, the numerous standards of the International Telecommunication Union (ITU) and the Moving Pictures Experts Group (MPEG, now ISO/IEC JTC 1/SC 29) are a rich source of detailed material for both technology and quality assessment standards [48]–[50].

Coming back to Figure 3, in case of a central bridging or hybrid bridging system (e.g., as proposed in [47]), one or more network conferencing bridges take over the mixing of the individual signals or streams (e.g., [43]–[46]) and may apply further processing for the *Network Access* (e.g., packet error correction) as well as further *Signal Enhancement* (e.g., echo cancellation, noise reduction, de-reverberation, automatic gain control).

From a QoE perspective, the methods of mixing the signals including any additional signal or data processing is a relevant factor. This holds for both network conferencing bridges and client-side conferencing bridges. When comparing client-side and central bridging technologies, a typical QoE-relevant difference is that central bridging usually requires additional transcoding steps. In turn, client-side bridging may not need this, but requires higher computational power in the clients and better network connections for the multiple streams.

Finally, all sites and any network bridges are connected over a network, whose characteristics (e.g., bandwidth, round trip delay, queuing strategies of routers, used network protocols) may influence the transmission of the packet streams along the delivery chain. Obviously, this can impact QoE, e.g., when end-to-end transmission delays get too long for fluent conversations, or when packet losses occur during unreliable transport and media payload is being lost.

2) SIFs RELATED TO USER-SYSTEM INTERACTION

In the previous section, the focus was on the exchange of information between participants over the system. However, a further component is to consider the interaction of the participants with the telemeeting system to achieve this information exchange. This relates to the disciplines Human Computer Interaction, Usability Engineering, and User Experience Design. While many publications for research, teaching, and practise are available in these fields, these are only partly related to the scope of this paper on aspects that contribute to the QoE of telemeetings. Three relevant publications are, for example, [89], [102] and [90], which address complexity challenges such as possible information overload, interface design, and system structure. Moreover, in order to structure this vast field, we identified four different types of interaction with a telemeeting system or behavior when using it: setting up the system, interacting with the system's user interface during a telemeeting, choosing a communication medium, and adapting the user behavior to the system characteristics. With respect to the different categories of QIFs, the first three types of interaction can be approached from a *System Influence Factors* perspective. In that respect, the next sections IV-A2.a to IV-A2.c refer to the following sub-categories of SIFs according to Table 2: Operational aspects - setting up a telemeeting, Operational aspects - controlling an ongoing telemeeting, and Media richness aspects. The last type of interaction, adapting the user behavior to the system characteristics, relates more to the *Human Influence Factors* and will therefore be addressed later in Section IV-C3.

a: OPERATIONAL SIFs - SETTING UP THE SYSTEM

Classical telemeeting solutions such as telephone conference bridges or fixed high-end telepresence rooms are systems which are prepared and set up by experts beforehand. While such systems still are in use today, they are increasingly complemented and partly superseded by individually used set-ups. With such legacy systems, often the participants just needed to dial in (telephone bridge) or use some control interface (telepresence room) to start the connection, but they were hardly requested to set up and configure the system and the connections as such. Looking at typical state-of-the-art telemeeting solutions, however, the situation is quite different: Today, participants are often also tasked with setting up the system or at least parts of the system. For example, common software-based solutions allow to connect to the telemeeting using different devices such as (laptop) computer, tablet or mobile/smart phone and they allow to connect extra headsets, handsfree terminals, cameras, and screens. In such scenarios, the participants need to select the proper audio and video devices, check the settings both in the telemeeting application and the operating system of the device, adapt the volume of microphone and loudspeakers, choosing an appropriate local (wireless) network, etc. From a QoE perspective, this complexity of setting up and configuring the telemeeting system is highly relevant. This is particularly the case when the participants encounter any problems concerning the system setup, for example when this happens just before a telemeeting or when it happens often.

Next to such problem-oriented influences on QoE, today's configuration possibilities may also contribute to a positive QoE by empowering the user to do things on their own. On the one hand, modern telemeeting systems have automated so many technical steps that it is actually possible for non-experts to carry out the set-up on their own. On the other hand, once users acquired sufficient experience and practice with the system, as many people will likely have during the Covid-19 pandemic, it becomes easier to solve most problems on their own, or enables users to give advice to other participants. To the best of the authors' knowledge, there is no research work published on the impact of a telemeeting system's setup complexity on QoE, with the exception of the related work in [89]. Hence, future work is required to further investigate this aspect.

b: OPERATIONAL SIFs - INTERACTING WITH THE SYSTEM's USER INTERFACE

Modern software-based telemeeting systems support multiple features beyond audio and video communication, such as screen sharing, annotation features, text chat and/or the management of participants. Note that systems often differentiate between users who are hosting the telemeeting and those who are participating. The hosts usually have more possibilities to interact with the system than the other participants. For example, in some systems the host needs to give screen sharing permission to others, or the host can define, which additional features can be used. Thus, users of modern telemeeting systems are not only requested to set up the system before a telemeeting (see Section IV-A2.a), but often they are also required to control the system during the telemeeting.

For that reason, the service providers or application developers are faced with key questions from the domains of User Experience Design and Usability Engineering, such as: How to design the user interface of the telemeeting system in such a way that hedonic and pragmatic needs are fulfilled? Hassenzahl and Tractinsky [103] go beyond this focus on solving problems and needs and recommend to ''design for pleasure rather than absence of pain''.

With a focus on mobile phones and services, Park *et al.* [91] approached the topic of designing a good user experience from an analysis perspective. Based on a literature review, interviews and an observation study, they identified a comprehensive list of sub-elements of User Experience and grouped them into three categories: usability, affect, and user value. This list reflects the resulting effects rather than the causes, and consists of aspects such as simplicity, effectiveness, learnability, flexibility, etc. Further work could obtain more insights about how telemeeting aspects contribute to these items, and in turn to User Experience and QoE.

Concerning the link between *User Experience* and *Quality of Experience*, Wechsung and De Moore [104] discussed the general similarities and differences between these two concepts. A short characterization of both concepts in form of a table can be found in the appendix of that publication, which is publicly accessible online [105].

Focusing on software applications running on mobile phones, among them also communication apps, Ickin *et al.* [106] obtained a number of insights on QIFs. Two of such factors were the performance and the user interface design of the applications. For the latter factor, the study participants reported issues such as locations and sizes of buttons, resizing and scrolling problems, or inefficient manual input. Ultimately, the choice of which application will be used in a given situation may be affected by such aspects, as well as the more communication- and media-transmission type characteristics.

c: SIFs RELATED TO THE CHOICE OF THE COMMUNICATION MEDIUM

As mentioned in the beginning of Section IV-A2.b, users of state-of-the-art telemeeting systems have the possibility to choose between different communication modalities: audio-only or audio with video, additional functions such as screen sharing, text chat, file transfer, joint document editing, etc. Moreover, users can also combine or switch between these modalities during the telemeeting.

Next to the user-interface-design perspective taken in the previous section, one can also look at the impact of this flexibility from a more contextual point of view: When, why, and how do participants select a specific one from those different communication modalities and features?

For such questions, concepts building on the Media Richness Theory could form a starting point. According to the theory proposed by Daft and Lengel [29], different types of media can be categorized by the richness of information they provide, for example with text being of less richness than video. This theory was originally developed in [29] with a focus on communication in management contexts, and it was developed at a time when many of today's communication features were far from being suitable for mass market introduction, either due to technological, societal, or financial reasons. Consequently, studies have revisited those concepts over the years for newly emerged communication technologies and for different contexts. In [30], for example, it was concluded that remote working teams would actually benefit from being able to select between differently rich media according to the tasks at hand and the people's cognitive styles (i.e., the way how they formulate and process concepts and information), as opposed to a general advantage of a ''higher'' media richness.

The discussion so far refers to situations in which individual participants are required to choose an appropriate communication channel. However, there are also situations in which it is not the task of the individual but of the telemeeting host to take this decision. Examples for such cases are virtual classroom scenarios, in which the teacher chooses the communication channel according to the didactic needs and permitted by the available resources. Other examples are virtual discussions or standardization meetings with a large number of participants, in which the meeting chair can opt to limit the communication channels upfront, e.g., to enforce a more formalized communication behavior of participants.

At first glance, the act of choosing a proper communication medium suggests that these considerations fall under the category *Mixed Influence Factors* (see Section IV-D and Table 5), which is true for aspects such as the user's knowledge about the system capabilities and limitations. One can also take a technology-driven perspective here, emphasising that a number of technical characteristics determine the media richness that the system is able to provide. Accordingly, such aspects are collected here as *Media Richness Aspects*, a sub-category of the *System Influence Factors*, see Table 2.

B. CONTEXT INFLUENCE FACTORS (CIFs)

Context Influence Factors (CIFs) refer to the contextual characteristics, more specifically to the physical, temporal, social, economic, task and any technical and information context, that influence QoE [2], [5]. With respect to telemeetings, CIFs essentially refer to the overall situation in which the telemeeting takes place. This means, not only the physical environments at the connected sites and temporal aspects play a role, but also the communication scenario and use case as such. In that respect, this section refers to all four sub-categories of *Context Influence Factors* in Table 3: Use Case, Communication Scenario, Communication Environment, Time Aspects.

TABLE 3. Overview of Context Influence Factors, organized along the sub-categories of Figure 2.

Category: Context Influence Factors

Sub-Category	Quality Influence Factor	Quality Relevance	Background
Use Case	Purpose of the telemeeting	[107]	[108]
	Needs and benefits	—	—
Communication Scenario	Number of participants	[27], [32]	—
	Number of sites	[27]	—
	Mixture of face-to-face and mediated conversation	[28, Chap. 1], [27]	—
Communication Environment	Acoustical situation	[109], [110]	[111], [112]
	Optical/lighting situation	[110]	[113]
Time Aspects	Temporal changes of the context, e.g., difference in usage time	—	[114]
	Time of the day for participants in different time zones	—	—

Notes: The column *Quality Relevance* cites either empirical studies directly investigating the factor's effects or publications discussing the relevance more from a theoretical point of view. Moreover, the relevance can refer either directly to QoE or to perception or communication aspects, which in turn are relevant for QoE. When no references are given, the factor is considered to be relevant from a practical perspective and requires further study. The column *Background* provides pointers to further literature.

It should be noted that this section touches only briefly upon the three latter types of factors, while a major part of this section concerns the use case and more specifically the topics of telemeeting purposes and collaborative working. The motivation for giving these topics more room is to provide a foundation for future work on a better understanding of the kind of situations in which a telemeeting is a suitable or even the most suitable choice of communication medium.

1) CIFs RELATED TO ENVIRONMENTAL AND TEMPORAL CONTEXTS

In the field of standardized quality assessment, the physical context, i.e., the communication environment, is usually considered by defining and setting requirements for the acoustical and lighting situation to be met when conducting a quality assessment test, see e.g., [111]–[113]. Example studies that have investigated the impact of the acoustical and lighting situation on QoE are presented in [109], [110].

With respect to the temporal context, to the best of the authors' knowledge, little research has been conducted on the impact on QoE when participants are located in different time zones or when there are differences in the context due to different time-linked social uses and habits, e.g., when having a telemeeting on a weekend vs. weekday, or during a local festivity. However, there is some body of knowledge on the complex relation between temporal changes of the system characteristics and the QoE formation processes, see Section V-B4. These considerations address a different aspect of time. Another aspect regarding time is the conversation structure of participants during a meeting, which is discussed in Section V-A.

2) CIFs RELATED TO THE COMMUNICATION SCENARIO

Next to the communication environment and time-related aspects mentioned in the previous section, additional QIF refer to how many sites are connected, how many participants are situated at each site and how this would lead to possible mixtures between face-to-face and mediated conversations. These aspects have been taken into account as another sub-category of QIFs under the term *Communication Scenario*. However, the relevance of these aspects becomes more apparent when considering the communication processes in Section V-A, especially with respect to how a mixture between face-to-face and mediated conversation can influence the communication and in turn QoE. Another aspect is the relevance of recognizing the speakers in a telemeeting and being able to locate their specific position (see Section VI-D3).

3) CIFs RELATED TO THE TELEMEETING PURPOSE

In this paper, a telemeeting is considered to serve a certain set of purposes or goals. The QoE experienced by individual telemeeting participants is influenced by the participant's perception of the extent to which those purposes or goals could be reached. To encourage future exploitation of such knowledge, the network planning tool ITU-T Recommendation G.107 [115] is an example in which first considerations of purpose – at least indirectly – have been included: when it comes to the impact of delay, different network planning parameters are recommended, depending on whether the service is intended for scenarios in which high, medium or low sensitivities to delay can be expected.

a: CATEGORIZING TELEMEETING PURPOSES

One way to categorize possible telemeeting purposes is to differentiate them into accomplishing tasks, fulfilling social needs, and exchanging information. As there are many different possible tasks, work reported in the literature often uses McGrath's task circumplex [108] to further categorize group tasks. This model structures tasks into four categories

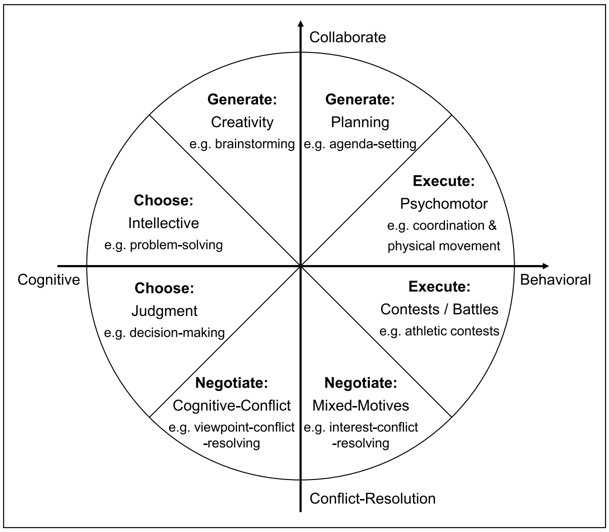

FIGURE 4. Two-dimensional circumplex model of group tasks – adapted from [108] and previously presented in [28].

(generate, execute, negotiate, and choose) along two dimensions (cognitive ⇔ behavioral, collaborate ⇔ conflict-resolution), see Figure 4. Examples of fulfilling social needs are telemeetings in which persons communicate to feel connected, to feel they belong to the same group, to get to know each other, etc. Finally, examples for exchanging information are making announcements, distributing news, or sharing useful information for group members.

A complementary way to categorize telemeeting purposes is to differentiate between professional / business telemeetings and private / leisure time telemeetings. Such a, sometimes non-binary, distinction can help to characterize telemeetings in relation to the conversation partners and their behavior, with aspects such as the degree of formality or expectations concerning the meeting outcomes. This approach is complementary to the one above that addresses the specific purpose. Both professional and leisure-time telemeetings can aim for accomplishing tasks and exchanging information, and also fulfilling social needs can play a role, not only in leisure time but also professional telemeetings, e.g., for improving commitment of individuals to a team.

b: DISTRIBUTED COLLABORATIVE WORK AS A FURTHER TELEMEETING PURPOSE

When a telemeeting serves the joint accomplishment of one or several tasks in a professional context, a common term to characterize such a telemeeting is remote or distributed collaborative working. This touches upon the multi-disciplinary research field of Computer-Supported Cooperative Work (CSCW), which Schmidt [116] characterizes as research to understand

> *cooperative work practices with the aim of contributing, both conceptually and technically, to the development of collaborative computing, i.e., computing technologies that facilitate, mediate, or regulate workers' interdependent activities.*

For an overview of main research threads in CSCW, the reader is referred to [117].

Focusing on distributed collaborative working using telemeeting technology, one important aspect for an effective and efficient collaboration is instantiating a shared workspace. This term refers to a physical or virtual space that allows the collaborating persons to share and jointly manipulate information and objects, e.g., see [118]. A typical example of a shared workspace in the physical domain is a meeting room with a whiteboard. For a telemeeting, a typical example of a state-of-the-art feature to create a virtual shared workspace is screen sharing that shows a virtual whiteboard or presentation slides.

As shared workspaces can take quite different forms, a first feature to characterize them is a differentiation between physical (or co-located) and virtual (or distributed) shared workspaces. Next to that, Park [119] proposed two more features: visibility, i.e., the extent to which an owner of information is sharing the view with the others, and controllability, i.e., the extent to which an owner of information is sharing the control with the others.

Nowadays, features such as screen sharing or joint document editing are commonplace examples for shared workspaces in many working contexts. With the advent of XR technologies, virtual workspaces can go beyond this, as they allow to (re-)create more immersive environments. Here, new questions arise when it comes to the combination of real and virtual environments as well as the potential benefits, which has been addressed by many researchers (e.g., in [120]–[122]). With current advances in remote sensing and control technologies in the area of cyber-physical systems, even more complex mixed physical/virtual shared workspaces are possible, in which physical objects can be manipulated by remote telemeeting participants.

Another aspect is the number of people who can simultaneously access such workspaces. Already with today's technological advances, this number has reached values way beyond 100 participants. Massive Open Online Courses (MOOC), virtual conferences, or virtual conventions are typical examples used in e-learning, academic, and business contexts. Another example, which in addition crosses the border from collaborative working to science entertainment, is the virtual telescope, which combined and processed data streams from multiple real-world telescopes to create a real-time virtual experience of a sun eclipse in June 2020 [123].

With respect to QoE, shared workspaces have an impact in two ways. First, the QoE experienced by telemeeting participants might include perceptual features and cognitive constructs regarding the shared workspace as such, e.g., in terms of video quality or system delay, or general usability. Second, the degree, to which the shared workspace actually supports the collaborative working process influences the participant's experience of that process, and in turn of the overall telemeeting.

c: FULFILLING SOCIAL NEEDS AS A FURTHER TELEMEETING PURPOSE

In this paper, social needs refer to the human desire to form and maintain social connections with other people. This relates to the feeling of belongingness to a group of people [124]–[126] as well as to the feeling of being connected with members of that group. To form and maintain such social connections, people want to communicate by expressing their views or sharing their knowledge with others, and by seeking information and opinions from others. Here, telemeetings and social media are two technologies that allow such communication with people situated at remote locations.

On one hand side, social media platforms have the potential to fulfill the need of belongingness and – to some extent – even the need of feeling of being connected. Recent studies investigate this potential but also possible drawbacks of social media and its relation to face-to-face contacts, see, e.g., [127]–[129]. On the other hand side, telemeetings allow for real-time and speech-based communication, which means that they have the potential to create an intense feeling of being connected. Based on the media richness theory [29], it could be assumed that telemeetings can create an even richer feeling of belongingness. Some support exists that perceived social belongingness is higher in face-to-face interactions than what can be achieved with text messaging [130]. Also, an underlying, pre-existing group belongingness for participants was reported to lead to a better QoE [131]. It is noted that the authors of this paper consider videoconferencing fatigue, often synonymously referred to as *Zoom Fatigue* in the recent literature, see e.g., [132], [133], as a constituent within a more holistic concept of QoE (cf. Section V-B). In turn, recent findings have challenged the assumption that videoconferencing may be preferred over text-based interaction, for the example in case of compensating for social distancing as required during the Covid-19 pandemic [134], [135].

Another aspect concerning the fulfillment of social needs by means of telemeeting technology is the feeling of co-presence [136], [137], i.e., the feeling of being there with the other person(s), or ‘‘a sense of being together in a shared space at the same time’’ [138], [139]. Another related term is that of social presence, i.e., ‘‘the sense of being together with a virtual or remotely located communication partner’’, which implies the feeling of co-presence and being in a communication with the other persons [138]–[141]. Here, a distinction may be made between group belongingness at large, and interpersonal bonds, where group belongingness may be achieved even with less rich information, while social presence in terms of interpersonal bonds can be increased by more face-to-face like cues, according to the work by [142] on distributed learning. A lot of research is ongoing in this area and can be expected to expand in the context of immersive media and Virtual Reality (VR) and Augmented Reality (AR) technologies, see e.g., [143]–[146].

C. HUMAN INFLUENCE FACTORS (HIFs)

According to [2], [5], Human Influence Factors (HIFs) refer to any characteristics of a user that have an influence on QoE, including the background and the mental, psychophysiological and physiological state of a user.

At first glance, HIFs refer to the person who is experiencing a multimedia system. When it comes to human-to-human communication over a telemeeting system, however, not only the HIFs of individual ‘‘experiencing person’’ are relevant, but also additional HIFs that relate to the other participants, their individual conversation behavior, as well as the relations between all participants. Based on these considerations, Table 4 provides a list of HIFs relevant in a telemeeting, grouped into the following subcategories: Characteristics of the perceptual and cognitive processes, Internal state of individual participants, Conversation behavior, Relations between participants, and Language aspects.

1) HIFs RELATED TO CHARACTERISTICS OF THE PERCEPTUAL AND COGNITIVE PROCESSES

HIFs strongly relate to the characteristics of the user's perceptual and cognitive processes [2], [5]. For example, impaired visual or hearing acuity will influence the perception of any degradations in the audio and video signals. When it comes to the list of HIFs in Table 4, the question arises to which level of detail these characteristics should be included. For example, there are many different possible forms of impaired visual acuity or hearing loss. Moreover, it is not clear in which way details about individual differences of the cognitive processing abilities of people, see, e.g., [150] can be taken into account either. For that reason, only the following, more global descriptors are used in this paper as HIFs: vision acuity, hearing acuity, olfactory acuity, tactile acuity, cognitive processing abilities. Here, the reader is also referred to Section V-B, which looks at the respective QoE-relevant perception and cognitive processes in more detail.

2) HIFs RELATED TO CONVERSATION PARTNERS AND CONVERSATION BEHAVIOR

One broad set of HIFs that are particularly relevant for a telemeeting refers to the participants, and more specifically to their communication goals and skills (including language and body language aspects), their individual mental state and personality as well as the relations between the different participants. These aspects influence the conversation behavior of the participants, for example regarding the amount of contributions of individuals, the way in which those contributions are made by the individuals and received by the other conversation partners, and the way in which the overall group conversation as such is managed. As the conversational behavior of participants influences the overall conversation structure and the communication processes (see Section V-A), the aspects discussed here are also relevant from a QoE

TABLE 4. Overview of Human Influence Factors, organized along the sub-categories of Figure 2.

Category: Human Influence Factors

Sub-Category	Quality Influence Factor	Quality Relevance	Background
Characteristics of the perceptual and cognitive processes	Vision acuity	[2], [5]	—
	Hearing acuity	[2], [5]	—
	Olfactory acuity	[2], [5], [147]	—
	Tactile acuity	[2], [5], [147]–[149]	—
	Cognitive processing abilities	[2], [5], [35]	[150]
Internal state of individual participants	Emotional state	[151]–[155]	[156]–[160]
	Stress	[161], [162]	[163]
	Attitude to subject	[23], [151], [164]–[166]	[167]
	Knowledge about subject	[168]	—
	Perceived importance of subject	[23], [168]	—
	Personal goals and intentions	[5], [169]	[170]
	Speaker's experience with lack of backchannel signals	[171]–[174]	—
	Propensity for immersion	[175]	[176]–[178]
	Susceptibility to simulator sickness	[179]–[182]	[183], [184]
Conversation behavior	Reactions to questions or statements	[107], [154]	—
	Amount of intended interruptions	[78], [79], [145], [185]	—
	Degree of moderation	[107]	—
	Formality of conversation	[107]	—
	Communication discipline in terms of conversation management	[107]	—
	Communication discipline in terms of content of contributions	[107]	—
	Active vs. non-active speakers / Degree of involvement	[186], [187]	—
	Pronunciation	—	[188]–[191]
	Listener's suppression of back channel signals	[192]	[193]
	Adaptability of communication behavior	[192], [194]	—
	Discipline to share information stemming from offline discussions	[107]	—
Relations between participants	Status	[107], [195]	—
	Role	[107]	—
	Trust	[196]	—
	Degree of mutual acquaintanceship / friendship / colleagueship	[197]	—
	Mixture of personal goals and intentions	[107], [169]	—
	Mutual expectations from each other	[107]	—
Language and body language aspects	Mixture of native and non-native speakers	[198]	
	Mixture of body languages	—	[173]

Notes: The column *Quality Relevance* cites either empirical studies directly investigating the factor's effects or publications discussing the relevance more from a theoretical point of view. Moreover, the relevance can refer either directly to QoE or to perception or communication aspects, which in turn are relevant for QoE. When no references are given, the factor is considered to be relevant from a practical perspective and requires further study. The column *Background* provides pointers to further literature.

point of view. One example is the finding that in certain conditions of transmission delay, active speakers are rated quality differently than passive listeners [186].

With respect to communication goals, the individuals' intentions and their positions in terms of knowledge and attitudes to the subject at hand influence the participants' communication behavior or the communication processes as such, see, e.g., [107]. In addition, the individual intentions can also influence the QoE formation process of that individual. More detailed discussions on this aspect are given, for instance, in [166] on the contribution of knowledge and attitude to the experiencing process and [151], [164] for observed links between attitude and QoE. Here, from an engineering perspective, it may be possible to infer the attitude from behavioral analysis, for example using conversation analysis, possibly even at a surface level, e.g., [199]. More details about the conversation process are given in Section V-A6.

With respect to communication skills, the individual's overall and momentary capabilities and willingness to cope with challenges of the discussion at hand (e.g., required cognitive load) as well as the system characteristics (e.g., lack of backchannels due to muted microphones) contribute in two ways. On the one hand, these aspects affect the participant's QoE as such, e.g., in terms of a discomfort due to a perceived lack of backchannels, when the participant is not used to it, or due to the impact of a required high cognitive load [32]. On the other hand, these aspects can influence the individual's communication behavior and thus also the experience of the other participants.

Similarly, the individual's internal state and personality are additional factors influencing communication behavior and QoE. The relation of emotion and communication is intensively discussed, for instance, in [155]. An impact of emotions or stress on QoE has been found for example in [151]–[153], [161]. With respect to personality, Schoenenberg *et al.* [197] found, for the case of transmission delays, that the personality that users perceived from other participants was linked to measures characterizing the conversation surface structure. Looking at personality from another perspective, Scott *et al.* [200] investigated the role of personality and cultural background on QoE. Obviously, if personality traits are perceived differently depending on the telemeeting system properties (e.g., [197]), it highlights the need for a holistic QoE assessment, beyond a mere audiovisual signal quality. If users do not use certain telemeeting platforms because the interaction with others is perceived as sub-optimal, even if not attributed to technology, the impact on technology acceptability will be just as bad as when the QoE-related issues are more explicitly attributed to the ‘‘communication channel’’.

Further aspects such as status and roles, trust, acquaintanceship and mutual expectations from each other as well as cultural aspects can determine the communication behavior between telemeeting participants. Here, studies on the automatic detection of roles, such as [195], [201] may be starting points for further analyses on the impact of roles on communication behavior and QoE. Finally, conversation management aspects such as moderation, agreed upon rules or degree of formality, are further factors. A framework for structuring the impact of roles and rules on conversation management is proposed in [107].

3) HIFs RELATED TO ADAPTING USER BEHAVIOR TO THE SYSTEM AND CONTEXT

Next to the considerations discussed above, there is a further type of participant's behavior in a telemeeting that is of particular interest from a methodological perspective: the users' tendency to adapt their behavior to the technical system characteristics and the context of the telemeeting. On the one hand, this refers to any adaptation of the conversation behavior depending on the system's capabilities and limitations as well as on the overall telemeeting context. On the other hand, this refers also to the topic of user-system-interaction, which was already mentioned in Section IV-A2.

From today's perspective, one general drawback of the Media Richness Theory mentioned in Section IV-A2.c is that it places face-to-face communication as the richest communication medium, which inherently means that face-to-face communication is the optimal way. This, however, is highly task or use-case dependent. While face-to-face meetings are definitely optimal for social interaction, they may be far less effective for decision making procedures or formal meetings. For instance, video access may impede the development of prosodic synchrony when some communicating partners display visually salient social cues, thereby dominating the conversation. In such conditions, communication via audio-only channels can be more effective in synchronizing speaking turns [202]. Over the years, several studies have shown that mediated collaboration can lead to similar or even better performance than face-to-face collaboration. This is for instance confirmed by a series of comprehensive literature reviews on decision support systems, which did not show a clear preference of face-to-face over mediated communication [14]–[17]. To account for such effects, Hantula *et al.* [31] proposed the Media Compensation Theory, which addresses the observation that humans actually adapt to electronic communication media; and Kock [203] proposed Media Naturalness Theory as a complementary approach by taking a behavioral perspective towards the use of electronic communication tools.

With respect to the topic of this paper, the degree to which participants are willing or able to such adaptation will influence QoE, both for them and for their conversation partners. This strongly relates to the individual's experience with the communication modality as well as the person's understanding of the system's capabilities and limitations. As an example of an effect on the participant: if the participant is not used to multiparty audio-only calls, that participant will experience a high cognitive load from the telemeeting, which in turn reduces the QoE. As an example of an effect on the others: if an inexperienced participant is too far away from a microphone to be adequately captured, the other participants will perceive a lower QoE, as the speech signal of that participant will sound degraded. In practice, communication between participants about such behavior- or usage-related problems often solves the issue, by accordingly adapting the technology usage.

D. MIXED INFLUENCE FACTORS (MIFs)

Next to the SIFs, CIFs and HIFs discussed in the previous sections, additional QIFs can be assigned to combinations of factors from the three main categories. Those factors refer to characteristics that are shared by two or all three of the main categories. An overview of those Mixed Influence Factors (MIFs) is given in Table 5.

Due to the large diversity of the MIFs, the following text discusses only a few examples that may be of particular interest, which are those factors that concern the interfaces between the physical environments at each site and the system. For more information about the remaining factors, the reader is referred to the references in Table 5 and to the supplementary material in [25], [26].

Looking at factors concerning the environment-system-interfaces, the first type of factors relates to the characteristics of the end devices: the signal transduction between the environment and the system, i.e., the electro-optical and electro-acoustical transduction, addressing, for example, the impact of background noise or ambient lighting. Further factors concern the extent to which a representation of the physical environments and of relevant objects in those environments as well as any communication-relevant side information is included in the transmitted signals.

These factors concern mainly characteristics of the system and the context. However, there is an additional group of factors concerning the environment-system-interfaces which also brings the human into the game: the positioning of the participants relative to the system components, and in particular to the capturing and reproduction devices. These types of MIFs are especially relevant from a QoE perspective. For example, non-ideal positions of speakers with respect to the microphones can lead to low QoE for the listeners, due to a reduced sound level, distance-induced coloration (due to the reduced level and high- and low-frequency audibility as well as the reduced direct-to-reverberant sound ratio), while optimal positions of viewers with respect to the displays can enhance QoE. Despite the QoE relevance of these positioning factors and their consideration in formal QoE test scenarios, see e.g., [23], [32], [74], [95], [113], [222]–[226], it is difficult to systematically address these factors in real-world settings. The reason is, that these factors are determined by a mixture of system, context and human aspects. This mixture could consist of limitations of the system, e.g., due to specific end devices used, constraints of the context, e.g., due to the interior of a room, and human behavior, e.g., with respect to the participant's awareness and willingness to change their position if that could improve overall QoE.

V. SURVEY ON QoE-RELEVANT PROCESSES CONCERNING TELEMEETINGS

After having discussed the large body research on QIFs, this section changes the perspective and looks at a number of communication, perceptual and cognitive QoE formation processes that are relevant for telemeetings. In that respect, it should be noted that this paper is touching on this field mainly from an engineering perspective and accordingly uses an engineering-type approach for describing the processes, e.g., by using flow diagrams. For that reason, this remark should be considered as a disclaimer in the sense that in other disciplines such as biology, neuroscience, psychology, or communication sciences, different descriptions are preferred.

A. COMMUNICATION PROCESSES

The primary purpose of a telemeeting is to communicate. Hence, the way in which the communication takes place is obviously a main contributor to the QoE perceived by the telemeeting participants. There is a vast amount of literature on human-to-human communication, both for face-to-face and mediated communication. In this paper, we focus on a number of aspects that have been considered in previous work with regard to QoE. First we present four inter-personal communication processes, i.e., processes that take place between the conversation partners: *Conversational Games*, *Grounding*, *Turn-taking*, and *Using Back-channel Signals*. After that, we discuss two further intra-personal communication processes: *Understanding* and *Response Formation*. Finally, this section closes with information on *Conversational Flow* and *Conversation Structure*; two concepts that help to characterize the degree of successful communication processes.

1) CONVERSATIONAL GAMES

Conversational games refer to parts of a communication that serve the accomplishment or alternatively the abandonment of a certain goal. Conversational games form a first step for separating a conversation into smaller units, as a conversation can consist of one to several conversational games. Conversational games have been introduced as a method to systematically characterize parts of a conversation with respect to the communication purpose, because they represent the *"pragmatic functions of utterances with respect to achieving speakers' goals"* [229]. More specifically, conversational games can be further separated into one or multiple conversational moves, i.e., utterances, which can be classified according to their purpose. In the literature, a number of coding schemes for conversational moves have been proposed [230]–[233], which were merged into a joint scheme in [28, Chapter 2]. To conclude, conversational games and moves allow the analysis of more complex conversations with multiple phases and even multiple communication purposes. Future work has to show, how this can also help in analyzing the QoE of telemeetings, which may be characterized by a set of complex conversations.

2) GROUNDING

Grounding [234] describes a process of establishing a mutual belief between speaker and listeners that an information has been correctly understood, i.e., that a common ground has been achieved. More specifically, Clark and Brennan [234] describe that this grounding process consists of a presentation

TABLE 5. Overview of Mixed Influence Factors, organized along the sub-categories of Figure 2.

Sub-Category	Quality Influence Factor	Quality Relevance	Background
Category: Mixed System-Context Influence Factors			
System Fitness for Use Case	Suitability of having a telemeeting instead of a face-to-face meeting (related to needs and benefits)	[204]	—
Signal transduction between environment and system	Electro-acoustical transduction (e.g., background noise)	[49], [205]	—
	Electro-optical/visual transduction (e.g., environmental lighting such as a too dark room or direct sunlight)	[206]	—
Handling of side information	Capturing, reproduction, creation of the (virtual) communication environment	[207], [208]	—
	Transmission of non-vocal communication signals	[172], [173], [209]	—
	Support of conversation management and turn taking processes	[210]	[9], [211]
Inclusion of physical environment	Inclusion of physical shared workspace	[120]–[122], [208]	—
	Cyber Physical System	[122], [212], [213]	—
Category: Mixed System-Human Influence Factors			
User's knowledge about the system	Technical expertise/affinity of participants with the system, its possibilities and limitations	[89], [214]–[218]	—
	Expertise to control the system, to fix problems	[89]	—
	Transparency of system responses to technical changes, e.g., change of IP address, bandwidth	[219]	—
	Degree to which user interaction is necessary to solve problems, e.g., re-establishing a dropped call	[89]	—
	Trust that the conversation is confidential (encryption, privacy, etc.)	—	—
	Operation of / interaction with microphones	—	—
Effects of delay	Perception of delayed responses from participants	[78], [79], [197]	—
Category: Mixed Context-Human Influence Factors			
Cultural aspects	Cultural background in terms of societal norms and expectations	[23], [220]	[221]
	Cultural background in terms of working styles	[23]	—
	Cultural background in terms of communication behavior	[23]	—
Category: Mixed System-Context-Human Influence Factors			
Position of participants in environment with respect to system components	Position of participants with respect to audio capture	[222] [95, Chap. 1.7]	—
	Position of participants with respect to video capture	[74]	—
	Position of participants with respect to audio reproduction	[32][95, Chap. 1.7]	—
	Position of participants with respect to video reproduction	[23], [113], [223]–[226]	—
Mental model of communication scenario	Mental model of conversation partner	[197]	[227]
	Focal assurance (certainty about who is talking)	[32]–[35]	[35]
	Mental model of the common (virtual) communication environment	[208], [228]	[227]

Notes: The column *Quality Relevance* cites either empirical studies directly investigating the factor's effects or publications discussing the relevance more from a theoretical point of view. Moreover, the relevance can refer either directly to QoE or to perception or communication aspects, which in turn are relevant for QoE. When no references are given, the factor is considered to be relevant from a practical perspective and requires further study. The column *Background* provides pointers to further literature.

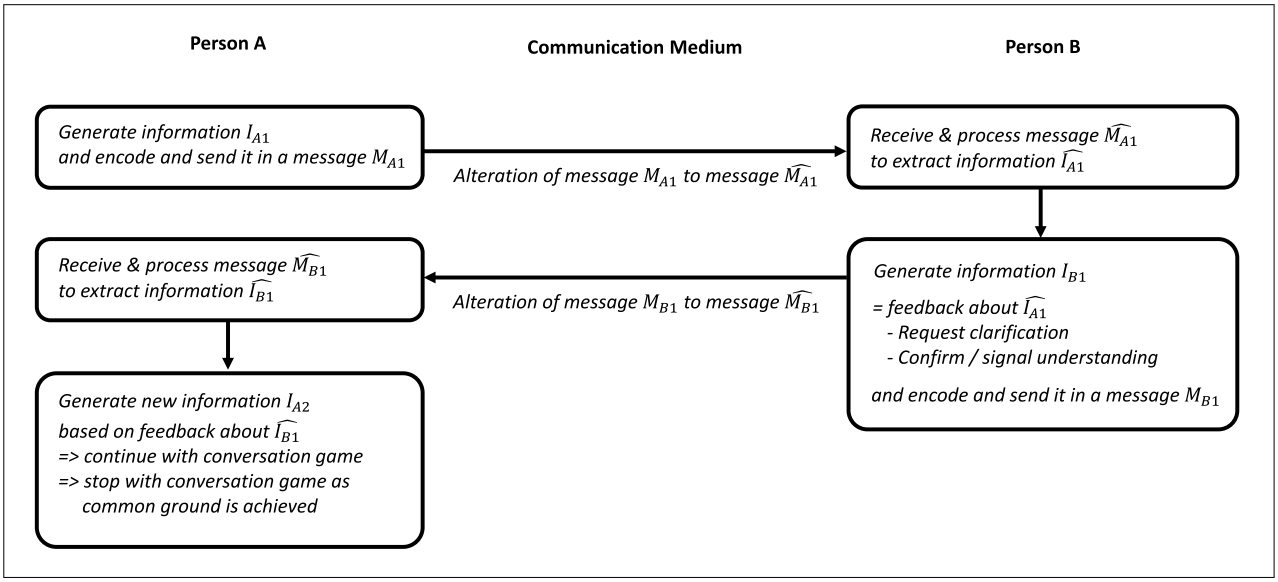
Person A
Communication Medium
Person B
Generate information I_{A1} and encode and send it in a message M_{A1}
Alteration of message M_{A1} to message $\widehat{M_{A1}}$
Receive & process message $\widehat{M_{A1}}$ to extract information $\widehat{I_{A1}}$
Generate information I_{B1}
= feedback about $\widehat{I_{A1}}$
- Request clarification
- Confirm / signal understanding
and encode and send it in a message M_{B1}
Alteration of message M_{B1} to message $\widehat{M_{B1}}$
Receive & process message $\widehat{M_{B1}}$ to extract information $\widehat{I_{B1}}$
Generate new information I_{A2}
based on feedback about $\widehat{I_{B1}}$
=> continue with conversation game
=> stop with conversation game as common ground is achieved

FIGURE 5. **The grounding process in the context of a mediated communication.**

phase (speaker's utterance) and an acceptance phase (listener feedback whether they understood the message or not). That means, it can take one or more turns until the grounding process for a particular message is completed. In the authors' view, the grounding process plays an important role in the user's QoE of a telemeeting. To start with, understanding the grounding process allows a quite analytic perspective on the potential impact of system characteristics on the conversation flow in a telemeeting. As Figure 5 explains in more detail, grounding in mediated communication may take a number of steps which can alter the original information that one person wants to convey up to the information that is actually understood by the other person. In the example shown in the figure, two persons A and B communicate, with their messages M_A and M_B in different phases of the process being altered, or the information I_A or I_B extracted from it. Those alterations can happen during the persons' perception, understanding and response formation processes as well as due to the communication medium. Here, it is the degree of such alterations that determines how much effort and how many turns participants need to spend for achieving the common ground: the stronger the alterations, the more effort and turns are necessary.

Following the argumentation in [234], it is commonly accepted that conversation partners usually have an intrinsic desire to reach common ground. Thus, any disturbance in the grounding process is assumed to have some negative impact on the perceived conversation and thus on QoE. In that regard, a disturbance can mean that the grounding process requires more effort than usual, or that the process as such is even temporally disrupted. Both technical and non-technical reasons can cause such disturbances. In turn, any means that ease the grounding process can increase QoE.

Finally, grounding is strongly connected with the other concepts considered in this section: the turn-taking process and back-channel signals described below are major components of the grounding process, while grounding with its specific purpose of reaching a mutual understanding can even be seen as one type of conversational games.

3) TURN-TAKING

Turn-taking refers to the transitions between speech utterances of each conversation partner, i.e., it describes who is speaking when, and how a change of speakers is accomplished. The fundamental principle of the turn-taking process is described in a model proposed by Sacks *et al.* [9]. With this pivotal work, Sacks *et al.* have made a foundational step to what today is called conversation analysis. This model considers speaker turns as a composition of turn construction units followed by transition-relevance places: Turn construction units are sentential, clausal, phrasal, and lexical constructions; one can understand them as units that carry information. The transition-relevance places are the moments at which a continuation of the current turn or a speaker change may occur; one can understand them as units that are used for signalling the temporal organization inside a conversation. Therefore, any impact on this process leads to an impact on the conversation flow that can be considered as a mediator to QoE. For instance, ITU-T Recommendation P.1305 [174, Sec. 8] describes how transmission delay can impact the turn-taking process: Consider that speaker changes may occur not only by explicit hand-over from the current speaker but also by self-selection from the listeners. Then, transmission delay can cause that listeners are "missing" the transition relevance place, which in turn disrupts the self-selection. Especially in a multiparty scenario, this can lead to severe false-start problems, meaning that multiple listeners attempt to get the turn, interrupt each other and need several attempts to sort out who can continue with the next turn.

4) USING BACKCHANNEL SIGNALS

Backchannels [172], also referred to as listener responses [173], are signals from the listener to the speaker to continue the turn. These signals can be produced vocally using verbal or non-verbal expressions, for example, utterances such as *"mm"*, *"uh uh"*, *"right"*, *"okay"*, and *"yes"*. Or these signals can also be sent by means of facial expressions, gestures, and posture changes, e.g., straightening the upper body part, head nods, and establishing mutual eye gaze. As these backchannel signals support the turn-taking process, obviously, the system's ability to transmit these signals can strongly influence the conversation flow. Moreover, the lack or degradation of such signals can cause a non-pleasant experience for a speaker in a telemeeting, as the following example from practise sketches: Especially in multiparty telemeetings, it is quite common that participants mute their microphones to avoid unnecessary noise. However, due to this absolute silence the speaker has no information, whether the other participants are still following or not, which for instance can cause feelings of uncertainty or which can even trigger the speaker to stop the turn and request any feedback. Notice that many state-of-the-art telemeeting systems provide signalling features such as hand raising or thumbs up, which can be considered as additional ways of sending backchannels to the speaker or initiate a turn.

5) UNDERSTANDING AND RESPONSE FORMATION

After discussing some essential interpersonal communication processes, this section focuses on communication processes within one person. In this paper, those processes are discussed in two main stages, *Understanding* and *Response Formation*. This links also to the considerations regarding the *Grounding* process sketched in Figure 5: to achieve a common ground, a listener needs to attempt to understand what the speaker was saying, and the listener needs to formulate some response to signal back whether the message is understood or not.

Having a closer look at *Understanding*, one can differentiate two levels of an achieved understanding: intelligibility, which refers to the understanding of the spoken words or full sentences from the acoustic signal, and comprehensibility, which refers to the understanding of the meaning in a larger, pragmatic application context. The degree to which an understanding in terms of intelligibility can be reached depends on numerous aspects, such as the listener's hearing capabilities, the listener's fluency of the language, the speaker's pronunciation and articulation, and the signal quality of the acoustical signal, which in turn is influenced by the system and the speaker's and listener's environments. Some work on the relation between intelligibility and quality has been reported in the literature, which will be discussed in Section VI-D2.

The degree to which an understanding in terms of Comprehensibility can be reached depends on the listener's world knowledge and in particular on the knowledge about the current topic domain and context. In addition, knowledge about the speaker and his or her intentions (*"What does he or she want to express or achieve when saying this?"*) helps to assess any consequences that can be drawn from the message, which is a crucial aspect of meaning extraction.

Moreover, language fluency of both speaker and listener can also strongly affect the degree of understanding. On the one hand, language fluency can help to improve intelligibility in case of degraded speech signals, as the listener can rely on his or her knowledge of the language in order to fill in gaps in the received speech signal, see e.g., [235]. One the other hand, language fluency can impact comprehensibility to such an extent that the perceived personality can be affected as well in certain contexts, see e.g., [236].

Having a closer look at *Response Formation*, the person not only takes the understood message into account but also other aspects. Further, world knowledge, and here in particular knowledge and assumptions about the speaker and any other telemeeting participants, will influence the content and form of the response (*"Is it fine to formulate a short response or is a longer explanation necessary? Is it fine to respond in a more direct and emotionally neutral manner or is it better to react in a more empathetic way?"*). In addition, the person's own intentions play a role as well.

Apparently, not only the person's world knowledge in general, but specifically the listener's knowledge and assumptions about the speaker and other telemeeting participants are important factors for the two processes *Understanding* and *Response Formation*. In the literature, this has been considered especially in the context of perspective-taking during the *Grounding* process, see e.g., [237]. This means, the degree to which a listener can recognize a speaker over the telemeeting system is an important aspect, see Section VI-D3.

To summarize, ensuring good *Intelligibility* and *Comprehensibility* are of paramount importance, as they not only determine the participants' QoE but are also crucial input for the participants' *Response Formation* and thus for their general communication behavior (see below). Therefore, these aspects deserve the maximum attention, in particular to understand the impact due to the sound devices (microphones, receivers, amplifiers) and transmission tools, but also to the speaker and listening environments.

6) CHARACTERIZING COMMUNICATION PROCESSES: CONVERSATION FLOW AND CONVERSATION STRUCTURE

Conversation flow refers to the efficiency and smoothness of the communication. In other words, the smoother and more efficient the communication processes *Conversational Games*, *Grounding*, *Turn-Taking*, and *Backchannels* are taking place, the better the conversation flow, and in turn the better the QoE. There are multiple aspects that can influence the conversation flow or one or more of the described communication processes. These aspects can stem from any of the three main categories of QIFs, for example when a speaker has a limited experience in coping with lacking *Backchannels* in terms of a HIF, a non-optimal mixture of face-to-face and mediated communication as a CIF, possibly mediated by technology and hence SIF, or a significant end-to-end transmission delay as a SIF.

Conversation structure can be analyzed at two different levels. On a first level, an analysis of conversation structure targets the components of a conversation in terms of their function during the communication process. When considering speech, such components are the individual utterances of the conversation partners. This perspective comes from the discipline of Conversation Analysis (e.g., see [9], [238], [239]), which builds on the analysis of turn-taking and repair processes (e.g., [9], [211]). As already mentioned above, the effect of disrupted turn-taking processes on telemeeting QoE is sketched in [174], for example regarding false start problems after interruptions due to transmission delay. However, future work is necessary to better understand the relation between QoE and conversation structure as a result of conversation analysis.

On a second level, the conversation structure can also be analyzed with regard to the sequence of on/off speech patterns in the conversation, irrespective of their function or content. This approach is also referred to as Conversational Surface Structure Analysis. Introduced in [240], [241], the principle is to describe the conversation structure as a temporal sequence of states in which no, one, or multiple speakers talk simultaneously. The advantage of this method is that – at least for speech-based analysis – it is rather straightforward to implement by means of voice activity detection algorithms. With this simplified analysis of conversations that does not require any speech recognition, such state-based surface structure models have also been investigated in a multi-modal analysis of conversations, e.g., [242]. Seen from a probabilistic perspective, Conversational Surface Structure is usually modelled as a Markov chain in which the steady-state and transition probabilities are obtained from observations. In [243], this approach has been used to characterize the effects of transmission delay on telephone conversations by computing statistical measures from a corresponding Markov model. Later, this approach has been further developed and extended with additional measures in the context of QoE evaluation of transmission delay [78], [79], [174], [185], [244]–[246]. Here, state probabilities and sojourn times, but also transitions between states at the different ends of a two- or multiparty communication can be used as sources of information, revealing, for example, unintended interruptions that may occur in case of delay, whether participants adapt their conversation behavior to delay, and whether the delay may be noticed as a QoE degradation and attributed to the system (cf. e.g., [78], [79], [199]).

B. QoE FORMATION PROCESS

Referring back to the definition of QoE in Section II-A and building on the fundamental work on quality perception in [6], it becomes clear that QoE happens largely in the user's mind. To better understand this perspective, this section takes a closer look at the processes inside the experiencing person. This leads to a more holistic understanding of telemeeting QoE, which in turn could help in technical system development. Note that in test contexts, the experiencing person is usually referred to as test subject, in real-life telemeetings that experiencing person is usually referred to as participant.

1) QoE-RELEVANT PROCESSES WITHIN THE EXPERIENCING PERSON

In the literature, a number of principles, taxonomies and models have been proposed to describe the formation of QoE or the link with related concepts such as Quality of Service, Quality Perception, and Quality Assessment, see e.g., [6]–[8], [92], [95], [247]–[250]. The motivation for such work is, for example, to provide insights on human quality perception that can help to improve technology – similar to approaches that exploit knowledge of human auditory or visual perception in coding – or to form the basis for instrumental quality assessment algorithms.

From an engineering perspective, there are two principal kinds of processes involved to form telemeeting QoE: those that process information, here referred to as *QoE-relevant Information Processing Mechanisms*; and those that steer the information processes, here referred to as *QoE-relevant Steering Mechanisms*. As the processing and steering mechanisms are tightly integrated in the human's mind, such a depiction only serves to illustrate the different components from a simplified, systems theoretic perspective. Figure 6 gives an overview of the main processes, which are described in the following.

First, let us focus on the *Information Processing Mechanisms*. Starting from the outcome of these mechanisms, *Telemeeting QoE* is the result of a *QoE Formation* process, which in turn consists of a number of sub-processes (details shown later in Section V-B2). In previous models (e.g., [8], [250]), this *QoE Formation* process takes as input the results of a *Perception* stage. Considering the state-of-the-art in perception research (e.g., [251]), this step is considered to be a pull mechanism (see also [252]). As a higher-level cognitive process, the *QoE Formation* process is taking the relevant information, such as the perceived characteristics of the audio and video signals, from a pool of *Perceptual Features*, with the pool in turn being filled by the underlying sensory perception processes.

As a novelty compared to earlier publications on QoE-formation processes, this paper explicitly considers another process as highly relevant for *QoE Formation* in a telemeeting context: *Communication*. Obviously, this *Communication* process is the most central cognitive process during a telemeeting, and it consists in itself of the two sub-processes *Understanding* and *Response Formation*, see the previous Section V-A. With respect to *QoE Formation*, the *Communication* process also takes *Perceptual Features* as input; however, the type and weight of information used by the *QoE Formation* and *Communication* processes may differ. Moreover, the *QoE Formation* process may also take as input some

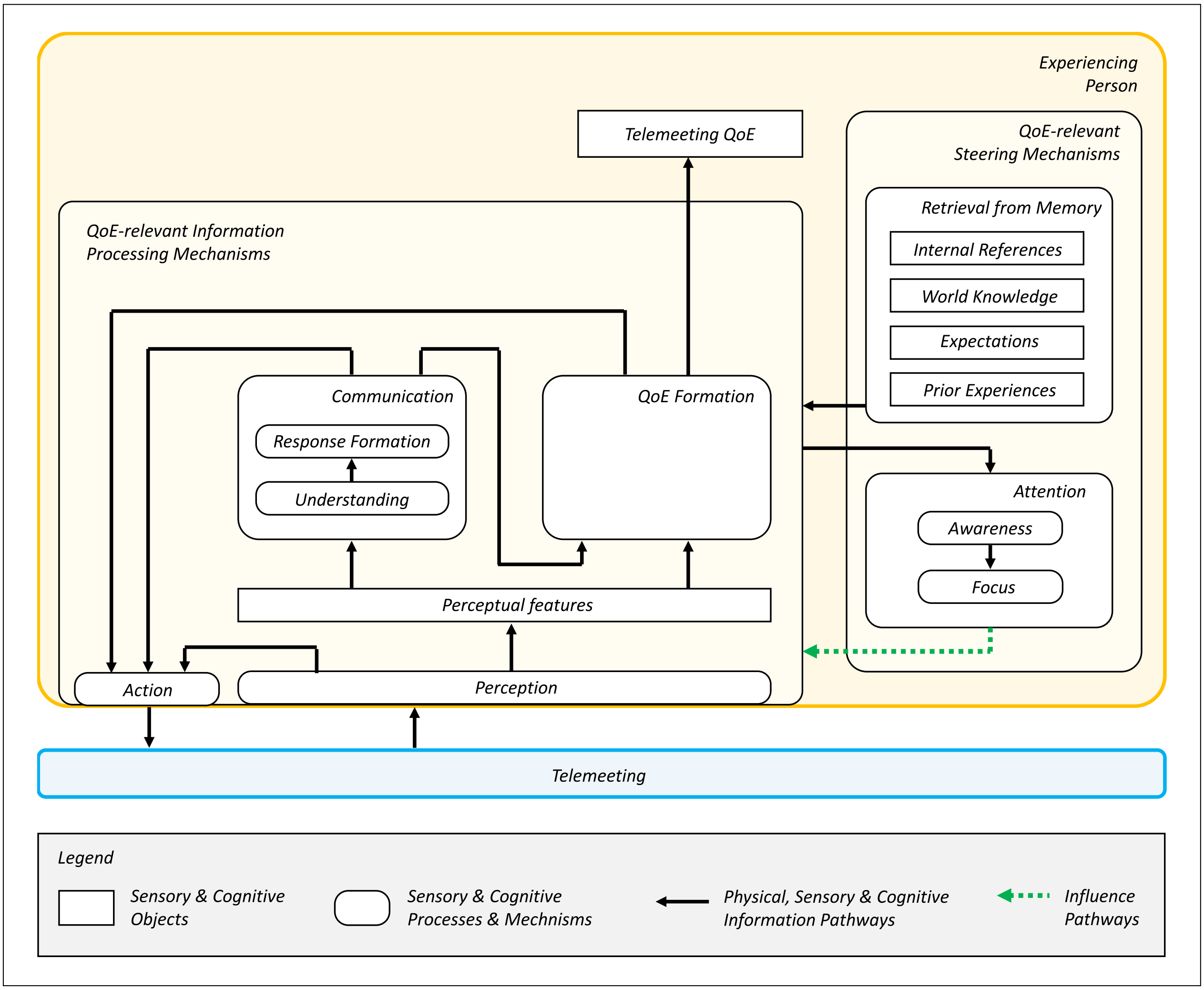

FIGURE 6. Overview of the QoE-relevant sensory and cognitive processes inside an experiencing person. See Section V-B1 for more details on the individual elements shown in the figure.

information specific to the *Communication* process, that is, the perceived characteristics from the *Perception* stage are augmented by further communication-related characteristics such as, for example, the conversation flow.

To complement the picture, an *Action* process is considered here as well. It accounts for the fact that an experiencing person performs different types of actions during a telemeeting: First, from perception research (e.g., [253]–[255]) it is known that people perform some actions to optimize the perception of a situation. An example for this is turning the head and eyes to the direction of a sound source to augment the auditory signal with visual information. Second, in a telemeeting a person obviously performs some actions to communicate: People speak and they also send other non-vocal communication signals (e.g., see [173]). Third, it has been observed that a certain QoE can trigger a person to different behaviors, see e.g., [248] for more details in video streaming contexts. This is also very common in telemeeting contexts; a typical example is that participants switch off video transmission when they experience quality problems with their network connection. Further, in case there are communication-related impairments such as for example background noise, interruptions in the audio channel, or a generally too low volume, users may speak up and apply Lombard speech (on the Lombard effect, see [256], [257]), ask a non-intelligible participant to repeat what she or he said, or increase the volume of their audio playout. For further listening-related measures see also [252].

Now, let us focus on the *Steering* mechanisms. The main process that can steer the previously mentioned *Information Processing Mechanisms* is *Attention*. Perception research has shown that attention is highly relevant during human information processing, as it enables people to tune the *Perception* process to individual signals, such as the speech signal of one particular conversation partner, resolving the cocktail party problem [258], [259]. This way, the mental information bandwidth and processing workload is effectively reduced. The bottom-up component of attention is typically referred to as saliency (see e.g., [260]–[262] for vision, listening and

QoE, respectively). In turn, attention can also be driven by top-down processes, for example when voluntarily attending to a specific conversation partner.

In the QoE domain, a similar impact can be asserted, as attention may be drawn or directed to individual QoE-relevant characteristics. For example, this may be the blurriness of a video signal from one particular conversation partner who is currently talking. Last but not least, *Attention* impacts the *Communication* processes as well, for example, by focusing and reacting on specific information.

Next to the aspect of *Focus*, *Attention* is also related to another sub-process: *Awareness*. The motivation for this is that attention needs a trigger. Such triggering may be based on information stemming from the *Perception* process: When the process fails to fully match the perceived, bottom-up sensory information with a number of top-down hypotheses [263], the person is becoming aware that something is missing or wrong, which in turn can trigger the person to pay attention and focus. Those mentioned hypotheses are stored in the person's memory and are referred to as internal references; the corresponding iterative sub-processes performing this hypothesis testing are referred to as *Anticipation & Matching* as a top-down process, and *Perceptual Event Formation* as a bottom-up process, see e.g., [8], [252], [28, Chap. 5].

Along with *Attention*, which is directly steering the different processes, further information is retrieved from the person's memory as additional input to the different processes. The *Internal References* already mentioned not only play a role during *Perception* but also in the *QoE Formation* process, when it comes to the formation of the desired quality features, see e.g., [6], [8], [28, Chap. 5]. Here, *Expectations* as well as *Prior Experiences* with the same telemeeting system or with previous similar telemeetings are taken into account. The terms *Internal References* and *Expectations* are often used in a similar way in a QoE assessment context. Internal references and expectations are assumed to result from *Prior Experiences*. According to Jekosch [6] and based on Piaget [264], both accommodation and assimilation may be involved in reference formation, depending on whether reference schemata are adjusted to the perceptual representation, or the perceptual representation to existing schemata, respectively. A more dedicated view on *QoE Formation* and *Expectations* can be found, for example, in [265]. For instance, *Expectations* may stem from other sources, such as costs or the particular situation, than from previous telemeetings, meaning that *Expectations* are not solely based on *Prior Experiences* and may instead be influenced by aspects such as advertisements, recommendations by friends, colleagues and family, or by reviews found on the internet or in magazines. Finally, during the *Communication* process, world knowledge enables the person to fully understand the perceived messages and to form appropriate responses.

2) QoE FORMATION PROCESS IN MORE DETAIL

According to Jekosch [6], at the core of *QoE Formation* is a cognitive process during which a quality judgement is formed by comparing the perceived features of an entity with the expected, desired features.

In the past, a number of extensions of this process model have been proposed to achieve several goals

a) to obtain a more detailed understanding of this *QoE Formation Process* and possible influences from inside and outside the person [8],
b) to explain a number of observed aspects, such as misattribution of technical quality problems to the interaction partners (see [7], [8] on process model extensions and [197] on the misattribution aspect),
c) to account for evidence indicating that QoE is essentially a multidimensional and multilayered construct (see below),
d) to provide theoretical models to explain this process in specific contexts such as the perception of asymmetries in multiparty meetings [28], [250] or when changing the viewing behavior in video streaming scenarios [248],
e) to embed this into broader characterization schemes of quality beyond the perception processes (see e.g., [92], [95], [247], [249] as well as Section VIII-A).

Building on those considerations, Figure 7 shows the essential details of this process. Contrary to previous work, however, the figure puts an emphasis on the fact that QoE is considered to be an aggregated construct of multiple, interrelated and time-dependent aspects, which here are introduced as *QoE Constituents*. Figure 7 visualizes this by separating the *QoE Formation* block into a stack of individual *QoE Constituent Formation* blocks, followed by an *Aggregation* stage. The green dotted arrows between the *QoE Constituent Formation* blocks indicate mutual interrelations.

In each *QoE Constituent Formation* process, further sub-processes are shown that reflect Jekosch's main principle combined with the most essential model extensions listed above. The input to each *QoE Constituent Formation* process stems from the *Perception* and *Communication* processes. In earlier work [8], [248], [250], we referred to this input as the perceived character, which consists of a multitude of *Perceptual Features*.

In a first step, those perceptual features are transformed into *Quality Features* by a *Reflection & Attribution* process. During the *Reflection* phase, only those features are selected from the totality of the *Perceptual Features* that are QoE-relevant. Here, the *Attention* process described before plays an important role. During the *Attribution* phase, any QoE-relevant features are either attributed to the telemeeting or to something else, such as, for example, the environment or the conversation partners. Here, mental models [227] of the telemeeting and especially the telemeeting system play an important role. During this stage also the *Desired Quality Features* are formed by retrieving information from memory, which reflect *Internal References* and *Expectations* in light of *Prior Experiences* of the person (see the considerations in Section V-B).

With the *Perceived* and *Desired Quality Features* as input, Jekosch's principle of *Comparison & Judgment* is the core

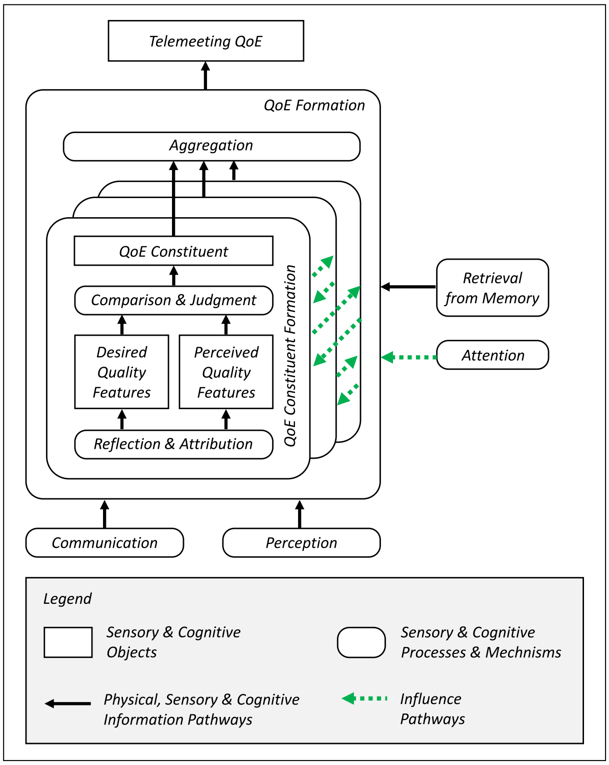

FIGURE 7. Visualization of the QoE formation process as described in more detail in Section V-B2.

stage, which forms a judgment about the *QoE Constituent*. Finally, the judgments about the individual *QoE Constituents* are aggregated to form an overall *Telemeeting QoE*.

Looking at this from an engineering perspective, the *QoE Formation* process can be considered as a sort of multi-dimensional signal processing mechanism: a multitude of input features are transformed, weighted, selected, compared, and aggregated to a multitude of intermediate representations and an output. Section V-B3 addresses this aspect of the multidimensionality of QoE in more detail, by discussing the relation between *Quality Features*, *Quality Dimensions*, and *QoE Constituents*.

Next to the multidimensionality, another perspective is to look at the temporal relations between the aspects discussed so far. On the one hand, the characteristics of a telemeeting can change during a meeting and thus also the *Quality Features*, *QoE Constituents*, and *Overall QoE* can vary. On the other hand, different *Quality Features* and *QoE Constituents* may be formed at different time scales and they may even influence each other over time. Section V-B4 provides more background information on the temporal aspects of QoE.

3) QUALITY FEATURES, QUALITY DIMENSIONS, AND QoE CONSTITUENTS AS PART OF THE QoE FORMATION PROCESS

In a typical approach taken in the literature – including our own prior work – an overall quality judgment is directly formed from a multitude of *Quality Features*, see top panel of Figure 8. As these features can be of quite different types, an approach has been presented in [278] to structure the features into four levels: level of direct perception, level of interaction, level of the usage scenario, and level of service.

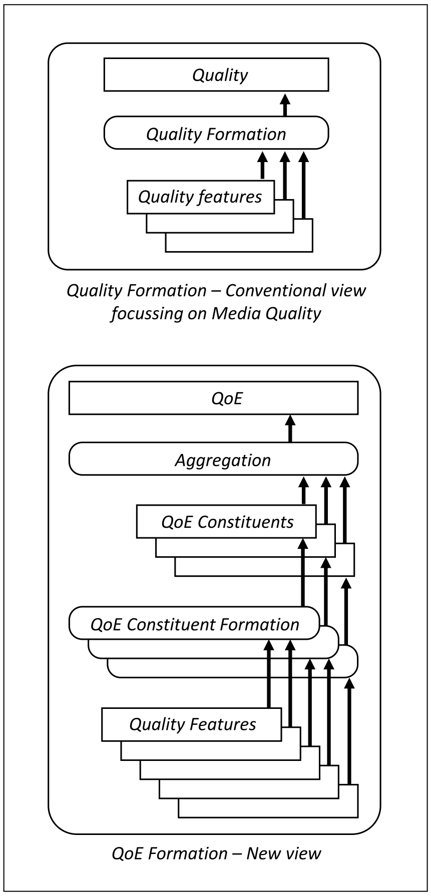

FIGURE 8. Visualization of QoE as a multidimensional construct that is formed by aggregating Quality Features and QoE Constituents. Top panel: conventional view in which a multitude of Quality Features are aggregated to an overall Quality. Bottom panel: extension with QoE constituents as an intermediate aggregation level. For details see Section V-B3.

TABLE 6. Overview of a number QoE Constituents that are highly relevant for telemeetings.

QoE Constituent	Quality Relevance	Background
Fatigue	[1], [37], [38], [133], [266]	—
Cognitive Load	[32]–[35]	[267], [268]
Immersion	[175]	[176]–[178]
Simulator sickness	[179]–[182]	[183], [184]
Feeling of Presence	[182], [269]–[272]	[175], [273], [139], [141]
Feeling of Co-Presence	[223], [274]–[276]	[273], [277], [137], [141]

Notes: The column *Quality Relevance* cites either empirical studies directly investigating the constituent in a QoE context or publications discussing the relevance more from a theoretical point of view. Moreover, the relevance can refer either directly to QoE or to perception or communication aspects, which in turn are relevant for QoE. The column *Background* provides pointers to further literature.

The present paper extends this concept by allowing for the formation of individual *QoE Constituents*, which then are aggregated to form an overall QoE judgment, see bottom panel of Figure 8. This extension essentially introduces an intermediate and visible level of aggregation: instead of directly aggregating a multitude of individual *Quality Features* into a single QoE judgment, the *Quality Features* are first aggregated into a set of *QoE Constituents*, which is then aggregated into an integral QoE.

The motivation for introducing these *QoE Constituents* is multi-fold. To start with, this still allows for the inclusion of research on multidimensional quality assessment, such as [51], [279]–[281], in which quality is considered to result from a set of orthogonal dimensions. These dimensions are extracted from a larger set of attributes and represent the underlying quality features. These dimensions may be integrated, for example, into audio or video quality, e.g., using preference mapping [279], [282]. Here, uni-modal media quality represents a QoE constituent.

Hence, the concept goes further than a solely dimension- and quality-based approach. First, different *QoE Constituents* need not be orthogonal since they may depend on common quality features. Second, *QoE Constituents* can encompass aspects that are not directly linked to speech, audio, or video signals, as it has so far been the focus of multidimensional quality assessment. Instead, also other *QoE Constituents* can now be considered in this framework, such as for instance simulator sickness, immersion, or fatigue, see Table 6.

4) TEMPORAL ASPECTS OF THE QoE FORMATION PROCESS AND THE NOTION OF QoE STREAMS

When it comes to temporal aspects of telemeeting QoE, the picture in Figure 9 is rather complex: the involved perception and cognition processes run in parallel, and there are different levels at which temporal changes can occur. On the first level, the technical and non-technical telemeeting characteristics are usually subject to changes over time. Audio and video signals per se are functions of time. Next, network and connection characteristics may change, and the system may respond to that with a certain behavior. In addition, participants may change their communication behavior; they may use different additional system features such as shared workspace or chat at different moments; or they may interact with the system interface a number of times during a telemeeting, etc.

On a second level, the participants' perception of the telemeeting is a function of time as well. At the level of perceptual processing, an example for temporal effects in auditory perception is temporal masking, see e.g., [283], [284]. At a higher level, auditory and visual objects and other perceptual features are formed based on the telemeeting characteristics captured by the human auditory and visual systems. Note that feedback mechanisms initiated at higher level may evoke top-down information that influences the bottom-up processing during auditory scene analysis [263]. However, there is not a strict one-to-one mapping of the temporal characteristics between the sensory input and the formed auditory and visual objects. For instance, in auditory perception research it is known that either single or multiple auditory objects can be formed from a multitude of short signal parts, and

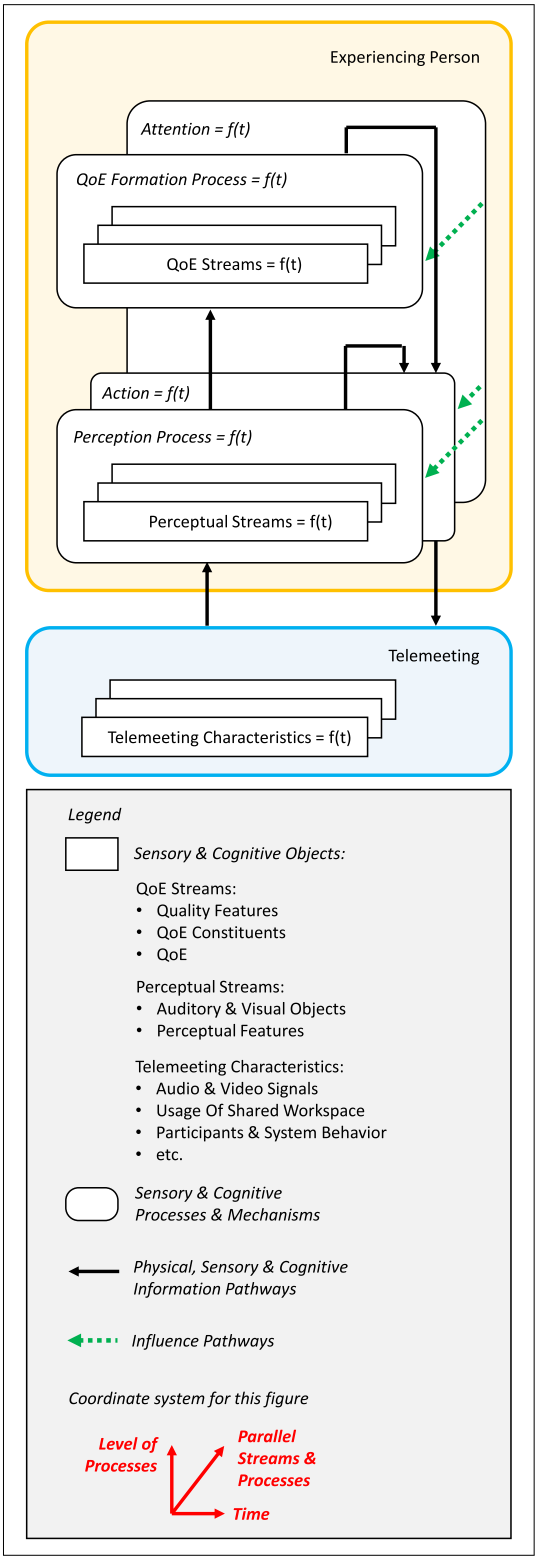
Experiencing Person
Attention = f(t)
QoE Formation Process = f(t)
QoE Streams = f(t)
Action = f(t)
Perception Process = f(t)
Perceptual Streams = f(t)
Telemeeting
Telemeeting Characteristics = f(t)
Legend
Sensory & Cognitive Objects:
QoE Streams:
• Quality Features
• QoE Constituents
• QoE
Perceptual Streams:
• Auditory & Visual Objects
• Perceptual Features
Telemeeting Characteristics:
• Audio & Video Signals
• Usage Of Shared Workspace
• Participants & System Behavior
• etc.
Sensory & Cognitive Processes & Mechanisms
Physical, Sensory & Cognitive Information Pathways
Influence Pathways
Coordinate system for this figure
Level of Processes
Parallel Streams & Processes
Time

FIGURE 9. Visualization of the temporal character of QoE formation processes and the notion of QoE Streams. For details see Section V-B4.

that this depends on the temporal and spectral characteristics of the acoustic input. This leads to the concept of auditory streams [285], or perceptual streams as a more general term, which allows to account for such temporal dependencies and effects. Next to these aspects, perception is also strongly influenced by attention, as discussed in Section V-B1. A person can change his or her attention focus between different perceptual streams at any moment in time.

On a third level, the *QoE Formation* processes are also a function of time. First, the formation of *Quality Features* and *QoE Constituents* (see Figures 7 and 8) is based on perceptual information, which is temporally changing. Thus the internal states and outputs of the *QoE Formation* process are time-dependent as well. Moreover, attention plays a role here, too, with users focusing on specific quality features at a time, or weighting these in a certain manner, see e.g., [8], [28, Chap. 5]. Hence, the *QoE Formation* processes as such can be influenced over time. In analogy to perceptual streams, this paper proposes to use the term *QoE Streams* when referring to the temporal evolution of the *QoE Formation* processes. For example, specific impairments identified in the audio or video signals of different participants may form such a *QoE Stream*, or the depiction of a screen share by one of the participants.

In addition to attention, action is another factor that contributes to the complexity, as the person's actions are functions of time that influence the perception and *QoE Formation* processes and vice-versa (e.g., [253]–[255]), as well as the telemeeting as such.

Next to such theoretical considerations, temporal aspects of QoE have also been empirically investigated for both momentary and episodic changes of signal quality, see e.g., [86], [286]–[288]. Moreover, some audiovisual quality models for non-communication-type media, such as ITU-T Rec. P.1203.3 [289], [290] for HTTP-based adaptive streaming contain specific considerations on temporal integration for the auditory and visual modalities. Similarly, the work in [291] has pointed to corresponding effects, where a base-quality was perceived by users when viewing audiovisual material at home, considering packet loss artefacts as additional impairments, as a sort of separate stream. Moreover, as this paper considers also other *QoE Constituents* than perceived signal quality, additional temporal aspects such as those of simulator sickness [184], presence [292], cognitive load and working memory, video conferencing fatigue [1], [131], [132], or usability and user experience, are also relevant for the formation of telemeeting QoE.

C. HOW QIFs AFFECT QoE FORMATION

It is obvious that the different steps of the QoE formation process may be influenced in different ways by the QIFs discussed in Section IV.

The most straightforward impact is that a QIF directly influences the sensory and cognitive processes within the experiencing person. For instance, a person could focus on certain *Quality Features*, when the person has a certain goal in mind, such as choosing a conferencing tool for a given purpose, or during a meeting, when she/he is in a certain emotional state. Or, the person might be distracted by events occurring as part of the context of use, e.g., during mobile use compared to stationary use in the office or at home. Or a person might not be very critical about the video quality because the person has some lower visual acuity, but is not wearing glasses or lenses. Another possibility is when a QIF has an impact on the telemeeting as such, for instance, in terms of achieving the meeting goals or having a good conversation flow, etc., which in turn has an impact on the perception of the telemeeting's QoE.

VI. SURVEY ON STATE-OF-THE-ART IN QoE EVALUATION OF TELEMEETINGS

In the following, the surveys on QIFs and communication and QoE formation processes are complemented by an overview of ''subjective'' and ''objective'' test methods for media quality and QoE evaluation. It is well known in the field that a QoE evaluation of a system is a nontrivial task, given the numerous QIFs that are relevant but not part of the system under test [326]. For that reason, a typical approach is to follow standardized test protocols to control such QIFs to a certain degree, or to explicitly include specific QIFs in the subsequent data analysis, as in the case of crowd-sourcing or outside-the-lab testing [327]. In this respect, the usage of standardized methods ideally ensures the reproducibility and comparability of the assessment results.

The next sections first provide a survey of the two main categories of available evaluation methods: (a) perceptual test methods, often referred to as subjective quality evaluation, and (b) instrumental methods, often referred to as objective quality evaluation. Then, some guidance is provided for the selection of a QoE assessment method that optimally matches the test case at hand. Finally, some complementary approaches to the QoE assessment of telemeeting systems are discussed.

A. PERCEPTUAL, SUBJECTIVE QUALITY EVALUATION

In perceptual tests, participants are invited to carry out, in a specific test context, certain tasks with the system under test. At certain times specified in the test standard, ratings of media quality or other measures related to QoE are collected, according to the specifics of the test protocol. In ITU-T Recommendations, for example, the test context, tasks, and methods to collect QoE-related ratings are often referred to as independent test factors, which can differ a lot between individual methods. Table 7 gives an exemplary overview of such aspects for a number of well-known perceptual test methods that are relevant for telemeetings. These methods can be considered as more conventional, direct perceptual test methods, as they ask test participants to give quality or other types of QoE-related ratings using a rating scale. The most prominent measures of quality obtained from such rating scales are Mean Opion Scores (MOS). For a precise definition of MOS and related terminology see [328].

TABLE 7. Overview of the main test factors for an exemplary set of standardized, perceptual (also referred to as subjective) QoE assessment methods, which are relevant for telemeeting systems or their components. Note: This is an updated version of a similar table presented in [28, Chap. 4]. Further, more recent methods specifically addressing telemeeting assessment are considered in the text, such as the P.1300 Recommendation series developed in Question Q10 of ITU-T Study Group 12.

Test Factor		**Perceptual (Subjective) Test Method**										
		ITU-T P.800 [111]	**ITU-T P.805** [293]	**ITU-T P.832** [294]	**ITU-T P.835** [295]	**ITU-T P.910** [113]	**ITU-T P.911** [296]	**ITU-T P.919** [297]	**ITU-T P.920** [298]	**ITU-R BS.1534** [299]	**ITU-R BS.1116** [112]	**ITU-R BT.500** [300]
Test Modality	Audio	✓	✓	✓	✓					✓	✓	
	Video					✓		✓				✓
	Audiovisual						✓	✓	✓			
Test Paradigm	Non-interactive (listening-only, viewing-only)	✓		✓	✓	✓	✓	✓		✓	✓	✓
	Conversation	✓	✓	✓					✓			
Presentation Mode	Single stimulus	✓	✓	✓		✓	✓	✓	✓			✓
	Multiple stimuli in temporal sequence	✓			✓	✓	✓	✓				✓
	Multiple stimuli presented simultaneously					✓				✓	✓	✓
Presentation of unimpaired reference	None	✓	✓			✓	✓	✓	✓			✓
	Hidden	✓				✓	✓	✓		✓	✓	✓
	Known	✓				✓		✓		✓	✓	✓
Rating Scale	Quality	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓
	(Specific) Impairment	✓	✓	✓	✓	✓		✓			✓	✓
	Other		✓	✓				✓				✓

B. INSTRUMENTAL, OBJECTIVE QUALITY EVALUATION

Contrary to the perceptual tests, instrumental evaluation approaches do not require the input from test participants to obtain a QoE rating about the system under test. Instead, instrumental approaches use an algorithm to predict media quality or other QoE-related aspects as they would have been rated by participants in a very specific test situation, and according to one of the previously mentioned perceptual test methods. Here, usually the average rating obtained from a group of participants is estimated, which is referred to as the Mean Opinion Score (MOS) [328]. The performance of such standardized QoE prediction models is usually validated in a rigorous manner within the standardization group, and in most cases based on validation test data previously unknown during model development. Nonetheless, such validation can be carried out only for a specific set of test factors according to the perceptual test methods that the prediction models are based upon. The underlying test methods, among other aspects, determine the modality of the predicted quality (speech, audio, video), and whether the quality prediction is for a noninteractive (listening- / viewing-only) or a conversation setting. Furthermore, instrumental approaches can differ in terms of input (ranging from system parameters for metadata models over bitstream information to the actual signals), and in terms of the usage of a reference for prediction (ranging from no- to reduced- to full-reference information being used, accordingly referring to the models as no-, reduced- or full-reference models). Table 8 gives an exemplary overview of QoE prediction models that are relevant for telemeetings. To the best of the authors' knowledge, those models have not been validated yet for different types of telemeetings and in particular not for multiparty settings, with the exception of Adel *et al.* [329], who investigated the performance of ITU-T Rec. G.107, the E-Model, [115] for codec tandems that occur in central-bridge-based telemeeting systems. Some concrete modelling ideas on how individual-channel model results could be employed for predicting a quality score for a complete multiparty meeting have been proposed in [330].

C. SELECTING APPROPRIATE EVALUATION METHODS

As a consequence of the different characteristics of the evaluation methods mentioned above, practitioners and researchers running a QoE assessment campaign need to opt for a test method that optimally matches the test case at hand. These test cases are often defined by system, processing and/or signal characteristics, as well as the use cases for which the telemeeting system under test has been designed. Next to the

TABLE 8. Overview of the main characteristics for an exemplary set of instrumental (also referred to as objective) QoE prediction models. Note: This is an updated version of a similar table presented in [28, Chap. 4].

Model Characteristic			QoE Prediction Models																				
			BS.1387 [301]	**G.107** [115], [302], [303]	**G.1070** [304]	**J.144** [305]	**J.246** [306]	**J.247** [307]	**J.341** [308]	**P.562** [309]	**P.563** [310]	**P.564** [311]	**P.862** [312], [313]	**P.863** [314]	**P.1201** [315]	**P.1202** [316]	**P.1204.3** [317]	**P.1204.4** [318]	**P.1204.5** [319]	**SSIM** [320]	**VMAF** [321]–[323]	**NISQA** [324]	**ViSQOL** [325]
Perceptual Test Protocol	BS.1116		✓																				
	BT.500					✓															✓		
	P.800			✓	✓					✓	✓	✓	✓	✓								✓	
	P.805			✓	✓					✓													
	P.910				✓		✓	✓	✓						✓	✓	✓	✓	✓				
	P.911														✓								
	P.920			✓																			
	Other																			✓			✓
Prediction Modality	Speech			✓						✓	✓	✓	✓	✓								✓	✓
	Audio		✓												✓								✓
	Video					✓	✓	✓	✓						✓	✓	✓	✓	✓	✓	✓		
	Audiovisual				✓										✓								
Prediction Paradigm	Non-interactive		✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	Conversation			✓	✓					✓													
Usage of Reference Signal	Full-Reference		✓			✓		✓	✓				✓	✓						✓	✓		✓
	Reduced-Reference						✓											✓					
	No-Reference			✓	✓					✓	✓	✓			✓	✓	✓		✓			✓	
Input	System Parameters			✓	✓												✓	✓	✓				
	Data	Packet Header										✓			✓								
		Packet Payload														✓	✓		✓				
	Signal	Raw Signal	✓			✓	✓	✓	✓	✓	✓		✓	✓				✓	✓	✓	✓	✓	✓
		Features					✓																
		Model-Internal Signal														✓							

Tables 7 and 8, several pointers are available that may be used for finding an appropriate QoE evaluation method. ITU-T Recommendations G.1011 [331], and especially P.1301 [27] and P.1310 [332] provide concrete guidance to QoE assessment methods suitable for telemeeting systems. In addition, the interested reader is referred to the overview pages of the corresponding ITU-T [48], [49], ITU-R [333], [334] and ISO MPEG [50] standards to get up-to-date information about standardized methods. Scientific texts such as [4], [20], [335] as well as text books and PhD theses on quality assessment of interactive telecommunication services (e.g., [7], [28], [51], [61], [82], [95], [246], [336] give further pointers to standardized and non-standardized methods that are relevant for telemeeting assessment. For surveys of QoE assessment methods for other services such as HTTP-based adaptive streaming, see, for example, [337]–[339] or generally for audiovisual multimedia, see [20].

D. ASSESSMENT APPROACHES BEYOND QUALITY RATINGS

1) CURRENT DEVELOPMENTS ON QoE ASSESSMENT METHODS

In the current state-of-the-art QoE assessment, additional aspects of QoE are increasingly moving into focus, and are assessed in different ways than using the conventional QoE-related, MOS-type rating scales. Highly relevant for telemeetings are approaches that look at the conversational structure, e.g., [174], cognitive load, e.g., [332], intelligibility of concurrent speakers, e.g., [340], and task performance e.g., [13]. Moreover, test methods for assessing 360° video QoE beyond media quality have been developed [341] and are provided in ITU-T Recommendation P.919 [297], addressing, for example, simulator sickness and viewing behavior.

Further approaches that help to assess the communication-related processes of Section V-A can be found in the large body of literature on intelligibility measurement and speaker recognition. The next two subsections outline the relation of these topics to QoE and provide pointers to relevant work.

2) SPEECH INTELLIGIBILITY: ASSESSMENT AND ITS RELATION TO QoE

Speech intelligibility is most commonly referring to word or utterance recognition in acoustic, verbal communication situations. The intelligibility of a spoken message depends on the speaker (e.g., articulation and speaking style) and listener (e.g., familiarity with the speaker's voice and the conversation context). Intelligibility is moreover influenced by the hearing abilities and the language proficiency of the listener. Intelligibility varies with the quality of the speech signal's acoustic transmission and the availability of visual cues from the speaker. Although a gold standard for speech intelligibility measurement is not available, there exist a number of standardized speech intelligibility assessment methods and models, see e.g. [342]–[345].

Concerning the relation between speech intelligibility and speech quality, a first approximation is that good intelligibility is a necessary – but not sufficient – prerequisite for good quality, see, e.g., [4]. This means, low speech intelligibility will result in low quality, but high speech intelligibility will not necessarily lead to high speech quality. Focusing on speech distortions induced by packet loss, Schiffner *et al.* [346] looked into this relation more deeply and showed a highly non-linear relation between intelligibility and quality: for rather high intelligibility, quality judgements can vary substantially but are hardly influenced by intelligibility, while for low intelligibility, quality judgements are consistently very low. Looking at background noise, speech bandwidths and speech levels, Preminger and Van Tasell [347] showed a similar complex relationship: In an experiment in which intelligibility varied between stimuli, the subjects hardly distinguished between the measured variables intelligibility, effort, and loudness.

The complex relationship between intelligibility and quality is also of high interest in the field of speech enhancement algorithms, both in telephony and hearing instrument contexts. It appears that algorithms can improve quality but not necessarily intelligibility, e.g., [348], or that not all algorithms that improve intelligibility also improve quality, e.g., [349].

3) SPEAKER RECOGNITION: ASSESSMENT AND ITS RELATION TO QoE

The importance for a listener to recognize the speaker's identity, to be able to associate specific opinions shared in a telemeeting to individual speakers, and to be able to form some impression about the speaker's personality has been investigated in different QoE-relevant contexts. In the context of grounding, Fussell and Benimoff [237] for instance discussed the importance of perspective-taking, in which the speaker's attempts to take the listeners' background knowledge into account facilitates comprehension. Looking at cognitive load and the underlying memory processes, Baldis [35] investigated the benefit of spatial audio reproduction on the listeners' degree of recognizing what each of the individual participants said, referred to as *Focal Assurance*. Other work re-evaluated this study, e.g., [34]; or picked up the aspects of cognitive load and focal assurance and investigated them in conjunction with complementary speech quality assessment questions [32], [33].

In terms of perceptual assessment methods, the studies cited above on cognitive load and focal assurance used both direct ratings and memory tasks, which were later also included in ITU-T Recommendation P.1310 [332]. In terms of objective assessment methods, a large body of literature is dedicated to the task of automatic speaker recognition and speaker identification, see, e.g., [350]–[356] for recent overviews. This body of methods can serve as basis for linking automatic speaker recognition with instrumental QoE assessment similar to [357]. This could complement existing work on direct, speaker-independent quality predictions from speech signals such as [324], [358], [359].

Another body of relevant work looks at the perception of personality using either perceptual or instrumental assessment methods. On the one hand, work investigated and predicted the link between personality traits and speech signals, e.g., [360]–[363], or the listeners' ability to recognize speakers over quality-impaired telecommunication channels, e.g., [357]. On the other hand, work investigated the link between perceived communication behavior, for example as a result of transmission delay, and the perceived personality, e.g., [197], [364].

VII. SURVEY ON CURRENT TRENDS CONCERNING TELEMEETINGS FROM A QoE PERSPECTIVE

After the overview of telemeeting QoE assessment methods presented in the previous section, this section provides more insights on the question in how far today's QoE assessment methods already cover near-future telemeeting systems. For that reason, this section looks at a number of relevant technological developments and, with a focus on XR-based telemeeting systems, it discusses a number of challenges concerning the QoE enabled by such systems as well as the corresponding QoE assessment methods.

A. FROM PLAIN OLD TELEMEETINGS TO eXtended REALITY (XR) & SOCIAL XR

Despite the progress in the past few decades, existing telemeeting solutions still have a number of drawbacks and restrictions that limit the users' communication experience. In this section, we revisit some of the QIFs discussed in Section IV from a technology development perspective. The goal is to discuss the aspects of QoE that near-future mediated communication solutions are likely to consider. As stated in [365], most

> *video conferencing tools [...] are geared toward voice-heavy, video-heavy, or PowerPoint-driven communications rather than collaboration.*

New developments in immersive communication in Virtual Reality (VR), Augmented Reality (AR), or Mixed Reality (MR) environments can close the gap in communication systems to allow more natural remote (computer-mediated) communication [366], as well as possibly allowing completely new forms of communication and interaction [367]. To evaluate such immersive remote communication, the authors consider social presence or co-presence as one of the key constituents to estimate the QoE of users, and as such how well an immersive telemeeting system can reproduce natural interactions [141].

The legacy videoconferencing systems discussed up to here in this paper have become a true alternative to physical meetings and traditional telephony. The usage of videoconferencing systems reached another level as a result of the Covid-19 pandemic during the years 2020 and 2021, when the world's population was forced to apply physical distancing as a strategy to fight the dissemination of the virus [368]. As a consequence, social presence and novel ways of mediated communication and virtual activities were sought more than ever before, see e.g., [369] on the tradeoff between physical and virtual activities.

Thus, with a strongly enhanced need for remote working and virtual get-together, there are many incentives throughout the telecommunication industry and research landscape to mitigate the drawbacks of current telemeeting solutions. Prolonged use of videoconferencing systems is found taxing on the HIFs (see Section IV) of the telemeeting QoE and may result in fatigue and increased cognitive load due to the unnatural communication setting, reduced mobility, and the additional effort required to send and receive non-verbal communication, an effect dubbed Zoom- or videoconferencing fatigue [1], [131]–[133], [266]. As several system influence factors mediate videoconferencing fatigue (see e.g., [1]), one solution is to improve the existing videoconferencing tools and streamline the communication experience (as discussed e.g., in [132]). Another direction is to create new solutions for the future that increase the naturalness and social presence as well as co-presence [370] – that is, the feeling of being in a place with one or more other persons at the same time – of mediated communication [138], [139]. These developments are aligned with recent advances in immersive technologies, in particular those enabling an eXtended Reality (XR) experience, that are promising to improve the immersion and presence in shared media consumption and communication [251]. This way, some of the interaction effects between QIFs related to the system (SIFs) and human users (HIF) of existing videoconferencing systems can be overcome. In order to successfully do so, the near-future XR systems must address a number of relevant SIFs, HIFs, and CIFs, reflecting the expected increase of data bandwidth, the need for real-time user tracking and novel system interfaces, for example.

Communication

- Business meeting
- Sprint / Stand-up / Scrum Meetings
- Negotiation
- Brainstorming
- Casual Co-Working
- Co-creation
- Social Co-presence (family, friends, ...)

Virtual Tour

- Architecture planning
- Virtual Guide (e.g., a tour in a museum)
- Sales presentation (i.e., presenting 3D models)

Conferences and events

- Large scale group communication
- Presentation
- Poster session
- Concert / sport event
- Lecture
- Networking
- Workshop

Remote Assistance

- Remote expert (both business / industry and private context)
- Doctor from a distance
- Call center (Ticket desk, help-desk)

FIGURE 10. List of most relevant uses cases for Social XR, partially based on [365], [371], [374].

XR is a term referring to all types of environments that employ Virtual, Augmented or Mixed Reality (VR/AR/MR) technology, and human-machine interactions enabled through computer technology and wearables. Here, the "X" represents a variable for any current or future spatial computing technology, or simply the "X" in *eXtended* Reality. One main differentiating factor for XR is the level of Degrees of Freedom (DoF) of user exploration and interaction, which expresses the level of freedom a user has to look at different parts and angles of the media content. The DoF goes from head-rotation-only 360-degree video (3-DoF, head movements in terms of pitch, yaw, and roll) to full movement as 6-DoF (3 DoF plus three translatory coordinates x, y, z) and offers different degrees of immersion (for more see e.g., [371]). When referring to XR systems that are designed for immersive communication, this paper refers to such technologies as Social XR, a term that is used both in industry, e.g., [372], and in science, e.g., [373].

B. STRUCTURED OVERVIEW OF CURRENT TRENDS IN SOCIAL XR

When looking at the past developments and current trends in Social XR, the authors observe two main types of target experiences and two main lines of technology development. In terms of target experiences, developments either aim for

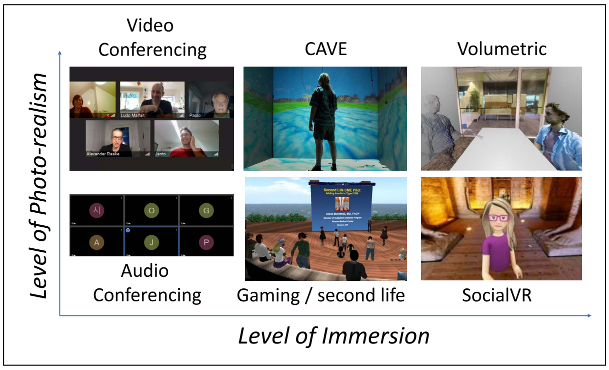

FIGURE 11. Matrix of communication technology (i.e., video conferencing, CAVE, Volumetric/Holograms, Second Life [375] and Social VR [376]) from two aspects: (1) level of immersion and (2) level of photo-realism.

getting XR as similar as possible to reality (i.e., extending or replicating reality) or aim for allowing experiences that are not possible in reality (e.g., by being partly or completely different by design, for example by enabling gaming-type functionalities such as being in certain, remote virtual places, or teleporting oneself between places). XR solutions may be applied for different telemeeting purposes, reflected in the CIFs such as the communication scenario and environment. Here, Social XR has a wide range of targeted and in part concretely specified use cases. Figure 10 provides a nonexclusive list of the most relevant use cases, partially based on [365], [366], [371], [374]. Several of these use cases substantially differ from each other. Therefore, it is currently hard to image a simple one-fits-all technical solution that will satisfy all requirements for all use cases. This can be illustrated when considering VR in comparison to AR, and corresponding differences and abilities with respect to rendering, the realization of meeting spaces with co-presence in different real or virtual environments, the enabled degrees of freedom in movement and interaction with users.

In terms of technology, developments either extend conventional video conferencing solutions [377] or aim for a completely new volumetric technology [377], [378], which brings them perhaps also closer to gaming technology. These technology development paths have profound effects on various SIFs, beginning from setting up and controlling the telemeeting to the aforementioned media richness aspects (see Section IV).

Another approach to characterize the technological evolution of communication systems towards Social XR is to look at the various improvements from two specific angles: Level of Realism, and Level of Immersion, as illustrated in Figure 11.

A high level of immersion at a low level of realism, i.e., based on computer graphics, is particularly impacted by principles of the computer gaming industry. One relevant development was Second Life (e.g., [375]) which offered a massive multiplayer immersive communication experience. With recent advances in VR Head-Mounted Display (HMD) technology, this resulted in several solutions to offer immersive VR experiences. An example of such VR communication platforms is Facebook Horizon, as the successor of Facebook Spaces [376]. With respect to the QIFs, increasing the level of immersion (for definitions, see, e.g., [139]) is intrinsically linked with one sub-category of HIFs, namely the Internal state of individual participants, feeling immersed and possibly present in a certain environment.

Looking at the increase in the level of immersion with a high level of realism, the existing, legacy computer-based videoconferencing services can be mentioned. One reason for their success is that these services aim for high video and audio quality, sometimes augmented with more advanced spatial audio capabilities, aiming to improve the media richness aspects, one sub-category of SIFs. Such tools are increasingly combined with further messaging and team-meeting capabilities, so that different teams may be created that can easily launch brief video-meetings if needed. Even more immersive videoconferencing solutions based on legacy technology exist, for example presenting visual information using a projection-based CAVE system (Cave Automatic Virtual Environment [379], or telepresence systems that use life-size displays spatially arranged around a common meeting table [380]–[382]. The intention of such systems is to increase interaction and more natural conversation by positioning projectors and screens to render users in life size.

For the future, setups such as Holoportation from Microsoft [383] promise to allow full body volumetric capture, transmission, and rendering of the user's body, usually referred to as holographic projection. As a consequence of such developments, different standardisation bodies are now starting new work items focusing on technical specifications for ''fully virtual meetings'' using holographic projection and aiming for setups including hotel halls, stadium, congress center, etc.

For these developments, placement and proxemics issues become important aspects of QoE. Placement refers to the relative location of different users and in particular to the question of how to place participants in XR so that they experience the same room when they are actually situated in highly dissimilar physical rooms/environments [384]. Proxemics refer to the requirement that users should respect each other's personal spaces, as the perceived interpersonal distance (proximity) is a significant determinant of social presence and quality of communication in immersive VR [385].

C. RENDERING TECHNOLOGY IN SOCIAL XR

One key component of Social XR, like any other XR application, is the rendering technology used. From a QoE perspective, the rendering technology determines a number of SIFs that can be considered in view of the media richness theory discussed in Section IV.

Rendering technology in the context of Social XR applications can be clustered along two dimensions. One dimension is the enabled DoF, ranging from 2D screens with the users confined to a rather narrow field of view in front of their screens [132], typically with non-spatial audio, over 360 VR (3-DoF) to 6-DoF VR or AR, including representations of the other participants that more plausibly integrate with the virtual (VR) or real environment (AR), including either head-tracked headphone- or loudspeaker-based spatial audio. The other dimension is the user representation, ranging from artificial avatars to photorealistic representations based on conventional video capture, or video- or geometry-based Point Clouds, or other volumetric representations [377], [378], [386].

For visual information, the term rendering defines the automatic process of generating digital images from three-dimensional models. A rendering engine can simulate an almost infinite and hence real-life-like range of illumination and color settings. However, current displays —- movie screens, computer monitors, etc. —- cannot handle the required peak luminance, contrast ranges and color gamut settings, so that some of the information must be discarded or compressed, reducing the resulting scene naturalness. Here, the fact that the human visual system also has its limits can help to suggest which short-cuts could be used in the rendering process to overcome technical limitations without a noticeable difference in user perception [387].

In addition to visual rendering, audio rendering in XR telemeetings may involve a dedicated positioning of the audio objects, representing the conversation partners at spatial locations that match the visually rendered scene an that therefore appear more natural to the user. A first step in this direction may be achieved by extending traditional telemeetings with spatial audio reproduction techniques [388]. Headphone playback without spatialization results in audio objects that appear localized inside the listener's head, or that all appear co-located in the same position, or with a reduced audio-visual spatial congruence of their perceived auditory and visual stimulus components. The QoE-related benefit of spatial audio rendering in telemeetings [32]–[35] as well as the spatial alignment of audio and video rendering [389], [390] have been investigated in the past. Professional telemeeting solutions with spatial audio have been introduced to the market as well, such as, e.g., Bluejeans [391], BT MeetMe [392] or the former Cisco telepresence solutions TX9000 [393] and IX5000 [394].

D. CHALLENGES IN SOCIAL XR

Despite the progress in Social XR, a number of technical challenges partially remain: In order to achieve perceptually plausible localization results with binaural headphone-based reproduction that includes appropriate externalization (e.g., [395]), head tracking or the usage of reasonably individualized head-related transfer functions become essential [396]. There are a number of further aspects that may become key factors in high-quality XR telemeetings, but are still open challenges, for example, simulating the acoustics of a virtual meeting room, embedding a virtual audio object in a real acoustic environment in an AR use case, removing the acoustic cues of the physical room where the speech signal is captured – see e.g., [395], [397], [398] on understanding such cues from a QoE perspective – and removing any unwanted background noise.

Besides an improved auditory scene analysis and support in solving the Cocktail Party problem [258], the interaction with others will also become more natural with spatial audio, beyond the spatialization and fixation to a 2D video screen [388]. Here, users can turn to others like they would, for example, during a face-to-face meeting or at a party, to engage in a temporal, smaller-group interaction.

Moreover, not only linguistic, but also nonverbal communication can be enhanced with XR-based technical mediation [399]: Facial expressions and especially bodily gestures can better be captured, transmitted, and displayed in XR. In the future, systems will likely even enable eye contact, like when being face-to-face in the same space. XR and a more holistic capture and display will also facilitate turn taking, since more cues indicating the intention to take the floor can be communicated.

With respect to user representation, a collection of rapidly developing technologies, including a suite of artificial intelligence (AI) tools, next-generation game engines, and augmented reality technology, bring on a new era of artificially intelligent avatars. An avatar, in this case, refers to any kind of user representation, either as a real person or as an artificial, simulated user agent. Directly related to artificial avatars is the *uncanny valley* effect [400]. The Uncanny valley describes an observation of human perception where a certain, yet imperfect level of human likeness of an avatar causes negative emotions and discomfort for users. While both a low and a very high level of human likeness are perceived positively, some level in-between is perceived negatively. Thus, the problem really starts when one combines or reproduces photorealistic representations of humans with computer-generated content. Interestingly, there also appears to be an ''Uncanny Valley of Telepresence'': the user's sense of telepresence (the illusion of ''being there'') increases with simulation quality up to a turning point, after which it begins to deteriorate, probably because the user's expectations start to exceed the actual affordances provided by the system [399].

Many technical advances have been made to unify XR platforms and devices (e.g., [401]), to capture virtual environments and users in photo-realistic quality, as well as to encode, store, and transmit 3D data (see [371]). Still, many technological limitations and challenges exist on each part of the XR ecosystem (i.e., XR frameworks, systems, and end-devices) [402]. Two particular new technologies that are expected to improve XR in mobile scenarios are 5G and remote rendering (in the cloud or at the network edge), especially as one can expect any XR device to be lightweight and thus potentially low-powered [381]. Both 5G and remote rendering individually and together will allow to shift resources from the end devices into the system, and thus to increase the rendering quality and performance of XR applications.

At this point it should be noted that this section was written as an initial overview of the technological trends and challenges of *XR*-based communication, and is by no means complete. A follow-up, in-depth paper will address this topic in more detail. In the present paper, the aforementioned concise technology review shall serve as the basis for the following, initial analysis of QoE assessment for XR-based communication.

E. QoE ASSESSMENT OF XR-BASED TELEMEETINGS AND SOCIAL XR

Understanding QoE in relation to Social XR is largely an open challenge. Partly, Social XR shares *QoE Constituents* with VR and AR, where, for example, measuring simulator sickness or quantifying the level of spatial presence have received a lot of attention [184], [403]. In the domain of AR and VR assessment, different assessment techniques are known, ranging from direct methods using questionnaires, e.g., regarding presence [175] or simulator sickness [183] to indirect methods using, for instance, physiological measurements [404], [405] or task performance, e.g., using wayfinding analysis [406], [407]. Furthermore, it is obvious that the underlying aspects of spatial auditory, visual and audiovisual perception and QoE evaluation play a role. Here, so far, only a few systematic or standardized assessment approaches exist. A set of aspects relevant in this regard is contained in the *Profile Template* instantiated in Section VIII-A.

Two other constituents, which Social XR brings to the attention of developers and researchers are co-presence, that is, the experience of being with others [137], [273] and social presence, that is, the feeling of co-presence and having an affective and intellectual connection with other persons [140], [141]. At the same time, a challenge for Social XR is the broader societal acceptance of being virtually and thus socially present in one location while being physically present in another, without being in contact with those in that physical environment. This might happen, for example, in the case of attending a virtual conference that may span over multiple days and occur in a different time zone, disrupting the daily routines of one's physically co-located social group, such as the family.

Lastly, an important challenge to consider is the ethics of XR use [408], [409]. XR telemeetings may make it easy to forget the rules of human interaction and enable immersive experiences that might be harmful or unpleasant. This can happen via inappropriate communication behavior due to cultural differences of the participants or due to the lack of physical co-presence and the resulting behavior mediation – see, e.g., [410] on rudeness in social media or [411] on rudeness in physical and computer-mediated work contexts. Or, this can happen due to errors and inappropriate decisions concerning the system design and development, see, e.g., [412] for existing industry guidelines on creating respectful, safe, inclusive, and accessible XR environments. One possible counter measure is the introduction of impenetrable personal zones to prevent that people can invade each others personal space, see, e.g., [413]. To summarize, an XR system, for which the level of realism can be controlled and content-induced risk is minimized, may be a key to a high-quality Social XR for all populations. For a review on these aspects, see [414].

Similar to the technological trends discussed before, this QoE-related section is intended as an entry point to the field of QoE assessment of XR-type telemeetings, indicating how prior work on VR and AR evaluation can form a basis for the case of interactive Social XR systems. A forward-looking analysis of telemeetings will be addressed in more detail based on the different projects and research activities running, for example, in the authors' different labs and institutions, and accompanying standardization activities in ITU-T Study Group 12 (Questions Q7, Q10 and Q13/12) and other standard development organizations (e.g., 3GPP SA4 IVAS project).

F. IMPACT AND FUTURE OF SOCIAL XR

First of all, it is important to stress that one should not regard Social XR as a replacement technology for any of the other existing communication channels. It is clear that telephone calls and traditional video conferencing will still have a clear value, at least in the near to mid-term future. However, with the further development of immersive communication and Social XR, many new use cases and a more natural interaction with high social presence will be possible [415], [416].

For the future, the authors expect that virtual and augmented reality and the real world can blend into each other and will completely change the way we experience mediated communication in general, and multiparty communication in particular. Here, XR communication has the potential to transform the everyday communication of people: On the one hand, by allowing new forms of communication in digital worlds that are currently not possible, and, on the other hand, by allowing better and more natural, intuitive communication between people. Technological breakthroughs towards more natural telemeetings might come in the form of understanding and modeling interaction intent (e.g., taking a holistic perspective of the Cocktail Party problem [258]), enhancing back-channel communication to disambiguate uncertainty (e.g., eye and face tracking to model gaze and facial expressions, posture tracking), novel interfaces beyond visual and auditory modalities, or adaptive and personalized communication systems based on user actions and feedback (e.g., individualizing audio delivery, correcting hearing or vision impairments).

This can have a direct impact on the life of tomorrow and may lead to a more sustainable future; to name a few benefits: inclusion of the elderly, inclusion of people with disabilities, reducing unnecessary travelling by providing adequate telemeeting alternatives, breaking communication barriers, or creating more awareness of world problems like diversity/climate change/populism, by virtually transporting people to the actual place of events, and enabling them to witness issues with their own eyes, concepts that would fall under the prospect of immersive journalism [417]. However, to elevate XR technology to this level, we need a much better understanding of the underlying user requirements and QoE in XR.

VIII. TOWARDS A HOLISTIC EVALUATION OF TELEMEETING QoE

Up to this stage, we have discussed the results of an extensive survey on the ingredients of telemeeting QoE, and have provided a short outlook on the future of telemeeting technology in the form of Social XR. One next, application-oriented step for a technical exploitation of this body of knowledge is to answer the question how these ingredients can be assessed in practice for a given telemeeting system. For that reason, this section builds on the survey on QoE assessment approaches by discussing a novel approach to characterize telemeetings from the holistic perspective endorsed in this paper.

A. PROFILE TEMPLATE FOR CHARACTERIZING TELEMEETINGS

Many of the more conventional QoE evaluation methods mentioned in Section VI are tailored to a specific test scenario, are focusing on certain individual aspects of a telecommunication system, or have not been developed with modern (multiparty) telemeeting systems in mind. To account for these drawbacks, existing efforts to guide investigators to an appropriate perceptual QoE evaluation method for telemeetings – and here in particular ITU-T Recommendations P.1301 [27] and P.1310 [332] – dissect the test cases at hand in order to identify the best matching existing evaluation method, as well as any potentially necessary adaptations of those methods. In that respect, those approaches already took the first steps towards a more holistic perspective on telemeeting QoE.

To attain a truly holistic perspective on telemeeting QoE, however, it is of great use to go one step further and provide a conceptual tool, which allows a systematic, agreed-upon and therefore comparable characterization of telemeetings to be obtained. This leads to the concept of a Telemeeting Profile Template, that is, a structured list of aspects that (a) characterize telemeetings and (b) are relevant from a QoE perspective. Benefits of such a characterization are, for instance, having a guidance when choosing an appropriate QoE assessment method, having a set of descriptors for a precise communication about telemeeting QoE, and having a means to develop a taxonomy of telemeetings. See Section VIII-C for further elaborations.

The Telemeeting Profile Template can be seen as a kind of check list containing attribute-value pairs. Accordingly, the Telemeeting Profile Template has two main columns: The first represents the list of characteristic aspects (the attributes); the second contains possible instantiations for each aspect (the values). As an example, one attribute in the first column is the communication modality, and the possible values are audio, visual, audiovisual, tactile, text-type, and graphics information for current and future, multisensory telemeeting systems [251], [418]–[420]. Apparently, combinations of values may be possible as well: Modern telemeeting systems allow to combine different communication modalities, e.g., audiovisual communication with additional text chat.

With respect to the list of identified attributes, the authors opted to refer to the Quality Influence Factors (QIFs), see Section IV, and use them as a tool to characterize telemeetings. As a consequence, this list of QIF-type attributes, was developed in the systematic way outlined in Section III,which went hand in hand with the literature survey for Section IV.

With respect to the values that are used for each attribute, one challenge is to find a good balance between covering all different possibilities and keeping the Telemeeting Profile Template manageable and comparable. This is especially the case when an attribute refers to some technology aspect which can actually have many different implementations. For that reason, more suitable values for technical aspects could represent a higher-level description instead of terms referring to variants of concrete implementations, such as *monotic*, *diotic*, *stereo*, *binaural*, *multichannel* as examples of values for the attribute spatial audio.

The resulting Telemeeting Profile Template is realized in the form of a large table which is provided in the supplementary material of this paper, i.e., a frozen, non-evolving version in [26] and a development version in [25], which can be modified, improved and extended also based on the feedback from readers of the present paper, as further outlined in the following Section VIII-B. To get a better overview of the information that constitutes the Telemeeting Profile Template, Table 9 provides an explanation about the different columns used in the supplementary material.

B. ONGOING DEVELOPMENTS CONCERNING THE TELEMEETING PROFILE TEMPLATE

This paper presents a first stabilized version of the Telemeeting Profile Template, more precisely the list of QIF-type attributes to be considered. The intention is to have a starting point for using and evaluating the Telemeeting Profile Template, in order to assess its validity and applicability. Moreover, additional work is necessary to obtain a set of concrete suggested values to complement the Telemeeting Profile Template. First suggestions by the authors can be found in the supplementary material of this paper [25], which is a commentable online document. Here, the authors plan to continue the development and invite interested researchers and practitioners to contribute.

Going one step further, having a standardized set of attributes and values will be ideal to address this challenge. Here, further work in research and practical application is expected to help improve the list of values, which eventually could even lead to a standardized list of recommended attributes and values. For that reason, the authors plan to continue refining the Telemeeting Profile Template, also based on readers' feedback, and to contribute a more stable version to ITU-T Study Group 12 for consideration as a future standard.

TABLE 9. Explanation of the columns in the Telemeeting Profile Template provided in the supplementary material, i.e., the frozen version in [26] and the development version in [25].

Column No.	Column Heading	Description
A – E	Columns corresponing to the tables on Quality Influence Factors in this paper	These columns contain essentially the same information as in Tables 3 to 5 in this paper.
A	Attribute Category	The highest hierarchy level used to structure the list of attributes. The categories are the different types of QIFs.
B	Attribute Subcategory	The second hierarchy level used to structure the list of attributes. A subcategory is defined by certain aspects that the attributes in that subcategory have in common. Those may be different between subcategories.
C	Attribute (i.e., Quality Influence Factor)	The individual attributes that constitute the Telemeeting Profile Template.
D	Scientific Quality Relevance	References to scientific publications that show the relevance of an attribute for the QoE of telemeetings.
E	Background	References to scientific publications that provide further relevant background information about an attribute or its context.
F	Examples of specifically concerned processes as described in this paper	In many cases, multiple of the QoE-relevant processes described in this paper are related to each individual QIF. This column only mentions example processes that are specifically interesting for or that help to understand an individual QIF. Note that this list of examples is for further study.
G	Suggested Values	Considering the Telemeeting Profile Template as a set of attribute-value-pairs, this column presents first suggestions for scale values that could be used. The scale values can be of nominal (identity only with itself), ordinal (values fulfill a certain order-relation), or interval character (values enable calculating additive properties). Note that this list of suggested values is for further study.
H – I	Scientific Relevance: More Detailed Info	These columns provide a more detailed discussion about the QoE relevance of Column D.
H	Studies reporting an influence on QoE or a relation between the attribute and QoE.	This column cites publications that show a rather direct link between the attribute and QoE. Additional comments in this column mention the essential aspect(s) of that link.
I	Other Scientific Evidence	This column cites publications that provide some supporting knowledge, although they may not show or discuss a clear proven link between the attributes and QoE. Additional comments in this column mention the supporting aspect(s).
J – L	Practical Relevance: More Detailed Info	These columns discuss the practical relevance of an attribute. This relevance was assessed by the authors with their different expertise in this field, while further feedback was given by experts from ITU-T Study Group 12 (Performance, QoS and QoE).
J	Direct Influence	This column discusses any direct influence of an attribute on QoE.
K	Indirect Influence	This column discusses any indirect influence of an attribute on QoE, meaning that the attribute is influencing a certain (set of) aspect(s), which in turn influence(s) QoE.
L	Example(s)	This column provides examples that further clarify the practical relevance of an attribute.
M	Definition, Description, Background Info	This column contains, when available, formal definitions of an attribute. Alternatively, this column contains descriptions by the authors or additional background information when appropriate.
N	Comments, Thoughts, Discussions	This column contains various comments, thoughts, and discussions that the author team was writing down when developing the Telemeeting Profile Template. In that respect, this column allows for the deepest glance into the development process of the Telemeeting Profile Template, which might help in a further development of the Telemeeting Profile Template.

C. WORKING WITH THE TELEMEETING PROFILE TEMPLATE

The individual deployment of the Telemeeting Profile Template depends on the actual use case. To illustrate the usage, the following paragraphs describe three possible application scenarios in more detail. The main target group considered in these three use cases are researchers and practitioners who are conducting QoE assessment campaigns of telemeeting systems, either during development or when the system is already in operation.

1) FINDING AN APPROPRIATE QoE ASSESSMENT METHOD

At this point, the Telemeeting Profile Template can help in two ways. On the one hand, a more systematic characterization of the telemeeting can assist to better specify the test scenario, which in turn helps to find an appropriate QoE assessment method more efficiently, using the pointers in the Telemeeting Profile Template as discussed in the previous paragraphs. On the other hand, such a detailed characterization of telemeetings can help to identify whether an existing method may be used without change, whether an existing method needs to be modified, or whether a new method needs to be developed.

2) COMMUNICATING ABOUT TELEMEETING QoE

Since telemeetings can be very different in their character, a precise communication about them can become challenging. This is, for instance, the case when a researcher or practitioner is asked to report about some QoE assessment campaign of a telemeeting system. Especially when a comparison with other systems is requested, the reporting person needs to be able to correctly interpret results in the context of the respective use case and system instances. Here, the Telemeeting Profile Template can help to characterize the respective telemeetings regarding the QIFs that have been addressed in the assessment campaign. This in turn minimizes the risk of misinterpretation and miscommunication.

One main challenge, however, is to balance between concise communication and using an extensive list of attributes. One possible approach is to separate between the analysis/comparison step and the communication step, that is, to consider the full template to identify all relevant commonalities or differences, while focussing on a set of main aspects in the communication. Here, future feedback from researchers and practitioners is sought to improve the usefulness of the Telemeeting Profile Template.

3) DEVELOPING A TAXONOMY OF TELEMEETING SYSTEMS

This use case picks up an underlying aspect of the two previous use cases: the potential benefit of a brief but precise categorization of telemeeting systems.

The purpose is to deploy the Telemeeting Profile Template in terms of a taxonomy. When selecting values that characterize different systems or the telemeetings typically held across these, the taxonomy-type character of the template becomes apparent.

One possible further work could hence be based on a data-driven approach: (1) characterize a representative number of different telemeetings using the Telemeeting Profile Template; (2) run a data analysis to identify a set of attributes that appear to be strong discriminators; (3) construct a first visualization of categories and the systems that belong to these, along with the set of attributes. As a result, the categories could be employed to analyze aspects such as system acceptance, user-groups typically employing these, or features that may be missing in specific cases. Another direction of future work is to systematically analyze mutual dependencies between attributes.

IX. CLOSING REMARKS

Telemeetings have been and will remain important for our professional and private lives, and are likely to become even more important in the future. This is shown by the developments during the past decades, the current situation during the Covid-19 pandemic, and recent technological, economic, societal, and climate-protection trends. Given such a major role of telemeetings, system developers and service providers should enable an optimal experience for the user, while at the same time keeping technical and financial resources at bay. For that reason, there is a need for understanding the detailed factors – ingredients – that contribute to a best possible *Quality of Experience* of telemeetings. And there is a need to be able to characterize those ingredients and their impact on QoE, both in a qualitative and a quantitative manner. Moreover, it is beneficial to understand in which directions the current technology developments are heading.

To address such needs, this paper analyzes current and near-future telemeeting services from a QoE perspective. In the first part of the paper, the authors provided an extensive survey of the numerous factors and processes that contribute to the QoE of telemeetings in order to achieve a holistic understanding of telemeeting QoE. As a next step, the paper introduces the current state-of-the-art of QoE assessment of telemeetings as well as ongoing developments. Then, the authors provided a glance towards the near future, where immersive technologies are considered to enable a new form of Social XR telemeetings with an improved experience of co-presence. Social XR will bring about new interaction interfaces, modalities, and types, which will require QoE evaluation methods beyond the current standards. To conclude the survey and technology outlook, the authors presented the Telemeeting Profile Template. It is a tool for practical guidance on telemeeting analysis and QoE assessment, intended to help with finding an appropriate QoE assessment method and creating a unified language for communicating about telemeeting QoE.

With the provision of a commentable, online version of the Telemeeting Profile Template [25], the authors wish to foster exchanges with other researchers and practitioners in the field, so as to expand the body of knowledge on telemeeting QoE assessment. In follow-up research and development work in their different institutions, the authors currently investigate how to best evaluate the QoE of future, Social-XR-type telemeetings, as will be described in corresponding future publications.

To wrap up, telemeetings represent a highly multidisciplinary field; they play an important role in our lives; they undergo promising technological developments; and they have the potential to provide global access to communication with other people, education, knowledge, and culture. The authors look forward to the upcoming forms of telemeetings in terms of technology, Quality of Experience, and usage scenarios and hope that this paper helps the interested reader to dive into a field that affects people and technology at the same time. The story of telemeetings – to be continued.

ACKNOWLEDGMENT

The authors would like to thank their colleague Elijs Dima and the anonymous reviewers of IEEE Access for their efforts to provide detailed feedback on this comprehensive paper. Moreover, they thank the colleagues from ITU-T Study Group 12 for their feedback on earlier versions of the Profile Template presented in this paper. The responsibility for the content of this publication remains with the authors. The International Audio Laboratories Erlangen are a joint institution of the Friedrich-Alexander-University Erlangen-Nürnberg (FAU) and Fraunhofer Institute for Integrated Circuits (IIS).

REFERENCES

[1] G. Fauville, M. Luo, A. C. M. Queiroz, J. N. Bailenson, and J. Hancock, "Nonverbal mechanisms predict zoom fatigue and explain why women experience higher levels than men," Tech. Rep., 2021. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3820035

[2] P. Le Callet, S. Möller, and A. Perkis, Eds., "European network on quality of experience in multimedia systems and services. Qualinet white paper on definitions of quality of experience, version 1.2," COST Action IC, Lausanne, Switzerland, White Paper, Mar. 2013. [Online]. Available: http://www.qualinet.eu/resources/qualinet-white-paper/

[3] D. Richards, *Telecommunication by Speech: The Transmission Performance of Telephone Networks*. Oxford, U.K.: Butterworths, 1973.

[4] S. Möller, W.-Y. Chan, N. Côté, T. H. Falk, A. Raake, and M. Wältermann, "Speech quality estimation," *IEEE Signal Process. Mag.*, vol. 28, no. 6, pp. 18–28, Nov. 2011.

[5] U. Reiter, K. Brunnström, K. de Moor, M.-C. Larabi, M. Pereira, A. Pinheiro, J. You, and A. Zgank, "Factors influencing quality of experience," in *Quality of Experience—Advanced Concepts, Applications, Methods*, S. Möller and A. Raake, Eds. Cham, Switzerland: Springer, 2014, pp. 55–72.

[6] U. Jekosch, *Voice and Speech Quality Perception—Assessment and Evaluation*. Berlin, Germany: Springer, 2005.

[7] A. Raake, *Speech Quality of VoIP—Assessment and Prediction*. Chichester, U.K.: Wiley, 2006.

[8] A. Raake and S. Egger, "Quality and quality of experience," in *Quality of Experience—Advanced Concepts, Applications, Methods*, S. Möller and A. Raake, Eds. Cham, Switzerland: Springer, 2014, pp. 11–34.

[9] H. Sacks, E. A. Schegloff, and G. Jefferson, "A simplest systematics for the organisation of turn-taking for conversation," *Language*, vol. 50, pp. 696–735, Dec. 1974.

[10] *Vocabulary and Effects of Transmission Parameters on Customer Opinion of Transmission Quality*, document ITU-T Rec. P.10/G.100, International Standard, International Telecommunication Union, Geneva, Switzerland, 2017.

[11] International Telecommunication Union. (2020). *Question 10/12—Conferencing and Telemeeting Assessment*. Accessed: Jun. 8, 2022. [Online]. Available: https://www.itu.int/net4/ITU-T/lists/q-text.aspx?Group=12&Period=17&QNo=10&Lang=en

[12] Lexico.com Online Dictionary, Dictionary.com and Oxford University Press. (2022). *Definition for: Telemeeting*. Accessed: Mar. 31, 2022. [Online]. Available: https://www.lexico.com/definition/telemeeting

[13] *Method for the Measurement of the Communication Effectiveness of Multiparty Telemeetings Using Task Performance*, document ITU-T Rec. P.1312, International Standard, International Telecommunication Union, Geneva, Switzerland, 2016.

[14] J. Fjermestad and S. R. Hiltz, "Experimental studies of group decision support systems: An assessment of variables studied and methodology," in *Proc. 30th Annu. Hawaii Int. Conf. Syst. Sci.*, vol. 2, 1997, pp. 45–65.

[15] J. Fjermestad and S. R. Hiltz, "An assessment of group support systems experimental research: Methodology and results," *J. Manage. Inf. Syst.*, vol. 15, no. 3, pp. 7–149, Dec. 1998.

[16] J. Fjermestad and S. R. Hiltz, "Case and field studies of group support systems: An empirical assessment," in *Proc. 33rd Annu. Hawaii Int. Conf. Syst. Sci.*, 2000, pp. 10–19.

[17] J. Fjermestad, "An analysis of communication mode in group support systems research," *Decis. Support Syst.*, vol. 37, no. 2, pp. 239–263, May 2004.

[18] B. Mennecke, J. Valacich, and B. Wheeler, "The effects of media and task on user performance: A test of the task-media fit hypothesis," *Group Decis. Negotiation*, vol. 9, pp. 507–529, Nov. 2000, doi: 10.1023/A:1008770106779.

[19] International Telecommunication Union. (2020). *Study Group 12 Performance, QoS and QoE*. Accessed: Jun. 8, 2022. [Online]. Available: https://www.itu.int/en/ITU-T/studygroups/2022-2024/12/Pages/default.aspx

[20] Z. Akhtar and T. H. Falk, "Audio-visual multimedia quality assessment: A comprehensive survey," *IEEE Access*, vol. 5, pp. 21090–21117, 2017, doi: 10.1109/ACCESS.2017.2750918.

[21] K. Bouraqia, E. Sabir, M. Sadik, and L. Ladid, "Quality of experience for streaming services: Measurements, challenges and insights," *IEEE Access*, vol. 8, pp. 13341–13361, 2020, doi: 10.1109/ACCESS.2020.2965099.

[22] M. Seufert, S. Egger, M. Slanina, T. Zinner, T. Hoßfeld, and P. Tran-Gia, "A survey on quality of experience of HTTP adaptive streaming," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 1, pp. 469–492, 1st Quart., 2015, doi: 10.1109/COMST.2014.2360940.

[23] J. B. Husić, S. Baraković, E. Cero, N. Slamnik, M. Oćuz, A. Dedović, and O. Zupčić, "Quality of experience for unified communications: A survey," *Int. J. Netw. Manage.*, vol. 30, no. 3, p. e2083, May 2020, doi: 10.1002/nem.2083.

[24] D. Vučić and L. Skorin-Kapov, "QoE assessment of mobile multiparty audiovisual telemeetings," *IEEE Access*, vol. 8, pp. 107669–107684, 2020, doi: 10.1109/ACCESS.2020.3000467.

[25] J. Skowronek, A. Raake, G. Berndtsson, O. Rummukainen, P. Usai, S. N. Gunkel, M. Johanson, E. A. Habets, L. Malfait, D. Lindero, and A. Toet. (2021). *Telemeeting Profile Template—Development Version. Hochschule für Technik—University of Applied Sciences Stuttgart*. Accessed: May 16, 2022. [Online]. Available: https://docs.google.com/spreadsheets/d/1uvr2ZRqEuW_xWBOi8UI8cUtRxb0qMJJChnHn5dDvYGU

[26] J. Skowronek, A. Raake, G. Berndtsson, O. S. Rummukainen, P. Usai, S. N. B. Gunkel, M. H. Johanson, A. P. Emanuël, L. Malfait, D. Lindero, and A. Toet, "Telemeeting profile template [data set]," Zenodo, 2022, doi: 10.5281/zenodo.6553448.

[27] *Subjective Quality Evaluation of Audio and Audiovisual Telemeetings*, document Rec. P.1301, International Standard, International Telecommunication Union, Geneva, Switzerland, 2017.

[28] J. Skowronek, "Quality of experience of multiparty conferencing and telemeeting systems—Methods and models for assessment and prediction," Ph.D. dissertation, Fac. IV Elect. Eng. Comput. Sci., Berlin Institute of Technology, Technische Univ. Berlin, Berlin, Germany, 2017, doi: 10.14279/depositonce-5811.

[29] R. L. Daft and R. H. Lengel, "Organizational information requirements, media richness and structural design," *Manage. Sci.*, vol. 32, no. 5, pp. 554–571, May 1986, doi: 10.1287/mnsc.32.5.554.

[30] M. Workman, W. Kahnweiler, and W. Bommer, "The effects of cognitive style and media richness on commitment to telework and virtual teams," *J. Vocational Behav.*, vol. 63, no. 2, pp. 199–219, Oct. 2003, doi: 10.1016/S0001-8791(03)00041-1.

[31] D. A. Hantula, N. Kock, J. P. D'Arcy, and D. M. DeRosa, "Media compensation theory: A Darwinian perspective on adaptation to electronic communication and collaboration," in *Evolutionary Psychology in the Business Sciences*, G. Saad, Ed. Berlin, Germany: Springer, 2011, pp. 339–363, doi: 10.1007/978-3-540-92784-6_13.

[32] J. Skowronek and A. Raake, "Assessment of cognitive load, speech communication quality and quality of experience for spatial and non-spatial audio conferencing calls," in *Proc. Speech Commun.*, 2015, pp. 154–175, doi: 10.1016/j.specom.2014.10.003.

[33] A. Raake, C. Schlegel, K. Hoeldtke, M. Geier, and J. Ahrens, "Listening and conversational quality of spatial audio conferencing," in *Proc. AES 40th Int. Conf.*, Tokyo, Japan, Oct. 2010, pp. 1–13, Paper 4–7.

[34] R. Kilgore, M. Chignell, and P. Smith, "Spatialized audioconferencing: What are the benefits?" in *Proc. Conf. Centre Adv. Stud. Collaborative Res. (CASCON)*, 2003, pp. 135–144.

[35] J. J. Baldis, "Effects of spatial audio on memory, comprehension, and preference during desktop conferences," in *Proc. ACM CHI Hum. Factors Comput. Syst. Conf.*, M. Beaudouin-Lafon and R. J. K. Jacob, Eds., Seattle, WA, USA, 2001, vol. 3, no. 1, pp. 166–173, doi: 10.1145/365024.365092.

[36] K. Kowalczyk, O. Thiergart, M. Taseska, G. Del Galdo, V. Pulkki, and E. A. P. Habets, "Parametric spatial sound processing: A flexible and efficient solution to sound scene acquisition, modification, and reproduction," *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 31–42, Mar. 2015, doi: 10.1109/MSP.2014.2369531.

[37] M. Urvoy, M. Barkowsky, and P. Le Callet, "How visual fatigue and discomfort impact 3D-TV quality of experience: A comprehensive review of technological, psychophysical, and psychological factors," *Ann. Telecommun.-Annales des Télécommun.*, vol. 68, nos. 11–12, pp. 641–655, Dec. 2013.

[38] M. Barkowsky, K. Brunnström, T. Ebrahimi, L. Karam, P. Lebreton, P. Le Callet, A. Perkis, A. Raake, M. Subedar, K. Wang, L. Xing, and J. You, "Subjective and objective visual quality assessment in the context of stereoscopic 3D-TV," in *3D-TV System With Depth-Image-Based Rendering: Architectures, Techniques and Challenges*, C. Zhu, Y. Zhao, L. Yu, and M. Tanimoto, Eds. New York, NY, USA: Springer, 2013, pp. 413–437, doi: 10.1007/978-1-4419-9964-1_14.

[39] W. A. S. Buxton, "Telepresence: Integrating shared task and person spaces," in *Proc. Graph. Interface*, Amsterdam, The Netherlands, Oct. 1992, pp. 123–129.

[40] M. Masoodian, "Human-to-human communication support for computer-based sharedworkspace collaboration," Ph.D. dissertation, Dept. Comput. Sci., Univ. Waikato, Hamilton, New Zealand, 1996.

[41] D. Jo, K. Kim, G. F. Welch, W. Jeon, Y. Kim, K.-H. Kim, and G. J. Kim, "The impact of avatar-owner visual similarity on body ownership in immersive virtual reality," in *Proc. 23rd ACM Symp. Virtual Reality Softw. Technol.* New York, NY, USA: Association for Computing Machinery, 2017, pp. 1–2, doi: 10.1145/3139131.3141214.

[42] D. Jo, K.-H. Kim, and G. J. Kim, "Effects of avatar and background types on users' co-presence and trust for mixed reality-based teleconference systems," in *Proc. 30th Conf. Comput. Animation Social Agents*, 2017, pp. 27–36.

[43] K. Singh, G. Nair, and H. Schulzrinne, "Centralized conferencing using sip," in *Proc. Internet Telephony Workshop*, vol. 7, 2001, pp. 57–63.

[44] P. J. Smith, P. Kabal, and R. Rabipour, "Speaker selection for tandem-free operation VoIP conference bridges," in *Proc. IEEE Workshop Speech Coding*, Tsukuba, Japan, Oct. 2002, pp. 120–122.

[45] P. J. Smith, P. Kabal, M. L. Blostein, and R. Rabipour, "Tandem-free VoIP conferencing: A bridge to next-generation networks," *IEEE Commun. Mag.*, vol. 41, no. 5, pp. 136–145, May 2003.

[46] S. Firestone, T. Ramalingam, and S. Fry, *Voice and Video Conferencing Fundamentals*. Indianapolis, IN, USA: Cisco Press, 2007.

[47] Y. Wu, C. Wu, B. Li, and F. C. M. Lau, "VSkyConf: Cloud-assisted multi-party mobile video conferencing," in *Proc. 2nd ACM SIGCOMM Workshop Mobile Cloud Comput. (MCC)*, 2013, pp. 33–38.

[48] International Telecommunication Union. (2020). *ITU-T Recommendations—G-Series: Transmission Systems and Media, Digital Systems and Networks*. Accessed: Mar. 31, 2022. [Online]. Available: https://www.itu.int/rec/T-REC-G/en

[49] International Telecommunication Union. (2020). *ITU-T Recommendations—P-Series: Terminals and Subjective and Objective Assessment Methods*. Accessed: Mar. 31, 2022. [Online]. Available: https://www.itu.int/rec/T-REC-P/en

[50] Moving Pictures Expert Group, International Organization for Standardization & International Electrotechnical Commission (ISO/IEC). (2020). *MPEG Standards*. Accessed: Mar. 31, 2022. [Online]. Available: http://mpeg.chiariglione.org/standards

[51] M. Wältermann, *Dimension-based Quality Modelling of Transmitted Speech*. Heidelberg, Germany: Springer, 2013.

[52] A. Spanias, T. Painter, and V. Atti, *Audio Signal Processing and Coding*. Hoboken, NJ, USA: Wiley, 2006.

[53] W. C. Chu, *Speech Coding Algorithms—Foundation and Evolution of Standardized Coders*. Hoboken, NJ, USA: Wiley, 2014.

[54] P. Vary and R. Martin, *Digital Speech Transmission—Enhancement, Coding and Error Concealment*. Chichester, U.K.: Wiley, 2006.

[55] S. Bruhn, H. Pobloth, M. Schnell, B. Grill, J. Gibbs, L. Miao, K. Järvinen, L. Laaksonen, N. Harada, N. Naka, S. Ragot, S. Proust, T. Sanda, I. Varga, C. Greer, M. Jelínek, M. Xie, and P. Usai, "Standardization of the new 3GPP EVS codec," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, South Brisbane, QLD, Australia, Apr. 2015, pp. 5703–5707, doi: 10.1109/ICASSP.2015.7179064.

[56] M. Dietz *et al.*, "Overview of the EVS codec architecture," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 5698–5702.

[57] Fraunhofer IIS. (2017). *The AAC-ELD Family for High Quality Communication Services*. [Online]. Available: https://www.iis.fraunhofer.de/content/dam/iis/de/doc/ame/wp/FraunhoferIIS_Technical-Paper_AAC-ELD-family.pdf

[58] P. A. Naylor and N. D. Gaubitch, Eds., *Speech Dereverberation*. Berlin, Germany: Springer, 2010.

[59] U. Zölzer, *Digital Audio Signal Processing*, 2nd ed. Chichester, U.K.: Wiley, 2006.

[60] J. Ramirez, J. M. Górriz, and J. C. Segura, "Voice activity detection—Fundamentals and speech recognition system robustness," in *Robust Speech Recognition and Understanding*, M. Grimm and K. Kroschel, Eds. Rijeka, Croatia: In-Tech Education and Publishing, 2007.

[61] B. Belmudez, *Audiovisual Quality Assessment and Prediction for Videotelephony*. Cham, Switzerland: Springer, 2014, doi: 10.1007/978-3-319-14166-4.

[62] M.-N. Garcia, *Parametric Packet-Based Audiovisual Quality Model for IPTV Services*. Cham, Switzerland: Springer, 2014.

[63] I. E. G. Richardson, *H.264 and MPEG-4 Video Compression—Video Coding for Next-Generation Multimedia*. Hoboken, NJ, USA: Wiley, 2003.

[64] T. Wiegand, G. J. Sullivan, G. Bjøntegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 7, pp. 560–576, Jul. 2003.

[65] L. Hanzo, P. J. Cherriman, and J. Streit, *Video Compression and Communications—From Basics to H.261, H.263, H.264, MPEG4 for DVB and HSDPA-Style Adaptive Turbo-Transceivers*. Chichester, U.K.: Wiley, 2007.

[66] *High Efficiency Video Coding*, document ITU-T Rec. H.265, International Standard, International Telecommunication Union, Geneva, Switzerland, 2019.

[67] *Versatile Video Coding*, document ITU-T Rec. H.266, International Standard, International Telecommunication Union, Geneva, Switzerland, 2020.

[68] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1649–1668, Dec. 2012, doi: 10.1109/TCSVT.2012.2221191.

[69] J. Nakamura, Ed., *Image Sensors and Signal Processing for Digital Still Cameras*. Boca Raton, FL, USA: Taylor & Francis, 2006.

[70] A. C. Bovik, *The Essential Guide to Image Processing*. Burlington, MA, USA: Academic, 2009.

[71] M. J. Tanakian, M. Rezaei, and F. Mohanna, "Digital video stabilization system by adaptive fuzzy Kalman filtering," *Inf. Syst. Telecommun.*, vol. 1, no. 4, pp. 223–232, 2013.

[72] P. Rawat and J. Singhai, "Review of motion estimation and video stabilization techniques for hand held mobile video," *Signal Image Process., Int. J.*, vol. 2, no. 2, pp. 159–168, Jun. 2011.

[73] G. Berndtsson, M. Schmitt, P. Hughes, J. Skowronek, K. Schoenenberg, and A. Raake, "Methods for human-centered evaluation of MediaSync in real-time communication," in *MediaSync: Handbook on Multimedia Synchronization*, M. Montagud, P. Cesar, F. Boronat, and J. Jansen, Eds. Cham, Switzerland: Springer, 2018, pp. 229–270, doi: 10.1007/978-3-319-65840-7_9.

[74] G. Berndtsson, M. Folkesson, and V. Kulyk, "Subjective quality assessment of video conferences and telemeetings," in *Proc. 19th Int. Packet Video Workshop (PV)*, Munich, Germany, May 2012, pp. 25–30, doi: 10.1109/PV.2012.6229740.

[75] *The Relative Timing of the Sound and Vision Components of a Television Signal*, document EBU Rec. R37, European Broadcasting Union, International Standard, Geneva, Switzerland, 2007.

[76] A. Vatakis and C. Spence, "Audiovisual synchrony perception for speech and music assessed using a temporal order judgment task," *Neurosci. Lett.*, vol. 393, no. 1, pp. 40–44, Jan. 2006, doi: 10.1016/j.neulet.2005.09.032.

[77] R. Eg, C. Griwodz, P. Halvorsen, and D. Behne, "Audiovisual robustness: Exploring perceptual tolerance to asynchrony and quality distortion," *Multimedia Tools Appl.*, vol. 74, no. 2, pp. 345–365, Jan. 2015.

[78] K. Schoenenberg, A. Raake, S. Egger, and R. Schatz, "On interaction behaviour in telephone conversations under transmission delay," *Speech Commun.*, vols. 63–64, pp. 1–14, Sep. 2014, doi: 10.1016/j.specom.2014.04.005.

[79] S. Egger, R. Schatz, and S. Scherer, "It takes two to tango—Assessing the impact of delay on conversational interactivity on perceived speech quality," in *Proc. 11th Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, Makuhari, Japan, 2010, pp. 1321–1324.

[80] N. Kitawaki and K. Itoh, "Pure delay effects on speech quality in telecommunications," *IEEE J. Sel. Areas Commun.*, vol. 9, no. 4, pp. 586–593, May 1991.

[81] T. Kurita, S. Lai, and N. Kitawaki, "Effects of transmission delay in audiovisual communication," *Electron. Commun. Jpn., I, Commun.*, vol. 77, no. 3, pp. 63–74, 1994.

[82] M. R. Schmitt, "Personal quality of experience: Accurately modelling quality of experience for multiparty desktop video-conferencing based on systems, context and user factors," Ph.D. dissertation, Vrije Universiteit, Amsterdam, The Netherlands, 2019.

[83] *One-Way Transmission Time*, document ITU-T Rec. G.114, International Standard, International Telecommunication Union, Geneva, Switzerland, 2003.

[84] S. Pierre, M. Barbeau, and E. Kranakis, *Ad-Hoc, Mobile, and Wireless Networks: Second International Conference, ADHOC-NOW 2003, Montreal, Canada, October 8–10, 2003, Proceedings*, vol. 2865. Berlin, Germany: Springer, 2003.

[85] V. R. Watzlaf and B. Ondich, ''VoIP for telerehabilitation: A pilot usability study for HIPAA compliance,'' *Int. J. Telerehabilitation*, vol. 4, no. 1, p. 25, 2012.

[86] D. Guse, ''Multi-episodic perceived quality of telecommunication services,'' Ph.D. dissertation, Fac. IV Elect. Eng. Comput. Sci., Berlin Inst. Technol., Technische Univ. Berlin, Berlin, Germany, 2016.

[87] S. Zielinski, F. Rumsey, and S. Bech, ''On some biases encountered in modern audio quality listening tests—A review,'' *J. Audio Eng. Soc.*, vol. 56, no. 6, pp. 427–451, 2008.

[88] S. O. Agnisarman, K. C. Madathil, K. Smith, A. Ashok, B. Welch, and J. T. McElligott, ''Lessons learned from the usability assessment of home-based telemedicine systems,'' *Appl. Ergonom.*, vol. 58, pp. 424–434, Jan. 2017, doi: 10.1016/j.apergo.2016.08.003.

[89] M. A. Sasse, M. J. Handley, and N. M. Ismail, ''Coping with complexity and interference: Design issues in multimedia conferencing systems,'' in *Design Issues in CSCW*. London, U.K.: Springer, 1994, pp. 179–195.

[90] O. Frick, ''Multimedia conferencing systems as building blocks for complex cooperative applications,'' in *Proc. Int. Workshop Multimedia Softw. Develop.*, 1996, pp. 61–68.

[91] J. Park, S. H. Han, H. K. Kim, Y. Cho, and W. Park, ''Developing elements of user experience for mobile phones and services: Survey, interview, and observation approaches,'' *Hum. Factors Ergonom. Manuf. Service Industries*, vol. 23, no. 4, pp. 279–293, Jul. 2013, doi: 10.1002/hfm.20316.

[92] S. Möller, K.-P. Engelbrecht, C. Kühnel, I. Wechsung, and B. Weiss, ''A taxonomy of quality of service and quality of experience of multimodal human-machine-interaction,'' in *Proc. Int. Workshop Qual. Multimedia Exper. (QoMEX)*, 2009, pp. –12.

[93] A. S. Chow and R. A. Croxton, ''A usability evaluation of academic virtual reference services,'' *College Res. Libraries*, vol. 75, no. 3, pp. 309–361, May 2014.

[94] S. Oviatt, ''Multimodal interfaces,'' in *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications, Second Edition*, J. A. Jacko, Ed. Boca Raton, FL, USA: CRC Press, 2012, pp. 405–430.

[95] S. Möller, *Assessment and Prediction of Speech Quality in Telecommunications*. Norwell, MA, USA: Kluwer, 2000.

[96] N. Côté and J. Berger, ''Speech communication,'' in *Quality of Experience*. Cham, Switzerland: Springer, 2014, pp. 165–177.

[97] M. Vaalgamaa and B. Belmudez, ''Audiovisual communication,'' in *Quality of Experience*. Cham, Switzerland: Springer, 2014, pp. 195–212.

[98] J. Skowronek, J. Herlinghaus, and A. Raake, ''Quality assessment of asymmetric multiparty telephone conferences: A systematic method from technical degradations to perceived impairments,'' in *Proc. 14th Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*. Lyon, France: International Speech Communication Association, Aug. 2013, pp. 2604–2608.

[99] V. Pulkki and M. Karjalainen, *Communication Acoustics: An Introduction to Speech, Audio and Psychoacoustics*. Chichester, U.K.: MIT Press, 2015.

[100] J. Lecomte, T. Vaillancourt, S. Bruhn, H. Sung, K. Peng, K. Kikuiri, B. Wang, S. Subasingha, and J. Faure, ''Packet-loss concealment technology advances in EVS,'' in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 5708–5712.

[101] W. Mack and E. A. P. Habets, ''Deep filtering: Signal extraction and reconstruction using complex time-frequency filters,'' *IEEE Signal Process. Lett.*, vol. 27, pp. 61–65, 2020.

[102] S. R. Hiltz and M. Turoff, ''Structuring computer-mediated communication systems to avoid information overload,'' *Commun. ACM*, vol. 28, no. 7, pp. 680–689, Jul. 1985, doi: 10.1145/3894.3895.

[103] M. Hassenzahl and N. Tractinsky, ''User experience-a research agenda,'' *Behav. Inf. Technol.*, vol. 25, no. 2, pp. 91–97, 2006.

[104] I. Wechsung and K. D. Moor, ''Quality of experience versus user experience,'' in *Quality of Experience—Advanced Concepts, Applications, Methods*, S. Möller and A. Raake, Eds. Cham, Switzerland: Springer, 2014, pp. 35–54, doi: 10.1007/978-3-319-02681-7_3.

[105] I. Wechsung and K. De Moor, ''Appendix to chapter 3: Quality of experience versus user experience,'' Springer, 2014. Accessed: Mar. 31, 2022. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-02681-7_3

[106] S. Ickin, K. Wac, M. Fiedler, L. Janowski, J.-H. Hong, and A. K. Dey, ''Factors influencing quality of experience of commonly used mobile applications,'' *IEEE Commun. Mag.*, vol. 50, no. 4, pp. 48–56, Apr. 2012.

[107] J. Mengis and M. J. Eppler, ''Understanding and managing conversations from a knowledge perspective: An analysis of the roles and rules of face-to-face conversations in organizations,'' *Org. Stud.*, vol. 29, no. 10, pp. 1287–1313, Oct. 2008, doi: 10.1177/0170840607086553.

[108] J. E. McGrath, *Groups: Interaction and Performance*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1984.

[109] T. Shimizu and H. Onaga, ''Study on acoustic improvements by sound-absorbing panels and acoustical quality assessment of teleconference systems,'' *Appl. Acoust.*, vol. 139, pp. 101–112, Oct. 2018, doi: 10.1016/j.apacoust.2018.04.021.

[110] S. Porcu, A. Floris, and L. Atzori, ''Towards the evaluation of the effects of ambient illumination and noise on quality of experience,'' in *Proc. 11th Int. Conf. Qual. Multimedia Exper. (QoMEX)*, Jun. 2019, pp. 1–6.

[111] *Methods for Subjective Determination of Transmission Quality*, document ITU-T Rec. P.800, International Standard, International Telecommunication Union, Geneva, Switzerland, 1996.

[112] *Methods for the Subjective Assessment of Small Impairments in Audio Systems*, document ITU-R Rec. BS.1116-3, International Standardization Report, International Telecommunication Union, Geneva, Switzerland, 2015.

[113] *Subjective Video Quality Assessment Methods for Multimedia Applications*, document ITU-T Rec. P.910, International Standard, International Telecommunication Union, Geneva, Switzerland, 2008.

[114] M. W. Howard and M. J. Kahana, ''A distributed representation of temporal context,'' *J. Math. Psychol.*, vol. 46, no. 3, pp. 269–299, Jun. 2002.

[115] *The E-Model: A Computational Model for Use in Transmission Planning*, document ITU-T Rec. G.107, International Standard, International Telecommunication Union, Geneva, Switzerland, 2015.

[116] K. Schmidt, ''Computer-supported cooperative work (CSCW),'' in *The International Encyclopedia of Communication Theory and Philosophy*. Atlanta, GA, USA: American Cancer Society, 2016, pp. 1–4, doi: 10.1002/9781118766804.wbiect144.

[117] K. Schmidt and L. Bannon, ''Constructing CSCW: The first quarter century,'' *Comput. Supported Cooperat. Work*, vol. 22, nos. 4–6, pp. 345–372, Aug. 2013.

[118] C. Gutwin and S. Greenberg, ''A framework of awareness for small groups in shared-workspace groupware,'' Dept. Comput. Sci., Univ. Saskatchewan, Saskatoon, SK, Canada, Tech. Rep. 99-1, 1999.

[119] K. S. Park, ''Enhancing cooperative work in amplified collaboration environments,'' Ph.D. dissertation, Graduate College, Univ. Illinois Chicago, Chicago, IL, USA, 2003.

[120] D. Reilly, S. Voida, M. McKeon, C. Le Dantec, J. Bunde-Pedersen, K. Edwards, E. D. Mynatt, A. Mazalek, and R. Want, ''Space matters: Physical-digital and physical-virtual codesign in inSpace,'' *IEEE Pervasive Comput.*, vol. 9, no. 3, pp. 54–63, Jul. 2010.

[121] T. Duval, T. T. H. Nguyen, C. Fleury, A. Chauffaut, G. Dumont, and V. Gouranton, ''Improving awareness for 3D virtual collaboration by embedding the features of users' physical environments and by augmenting interaction tools with cognitive feedback cues,'' *J. Multimodal User Interface*, vol. 8, no. 2, pp. 187–197, Jun. 2014.

[122] D. Clergeaud, J. S. Roo, M. Hachet, and P. Guitton, ''Towards seamless interaction between physical and virtual locations for asymmetric collaboration,'' in *Proc. 23rd ACM Symp. Virtual Reality Softw. Technol.* Gothenburg, Sweden: Association for Computing Machinery, Nov. 2017, pp. 1–4, doi: 10.1145/3139131.3139165.

[123] G. Masi. (2020). The virtual telescope. Bellatrix Astronomical Observatory, Italy. Accessed: Mar. 31, 2022. [Online]. Available: https://www.virtualtelescope.eu/2020/06/16/21-june-2020-solstice-annular-solar-eclipse-live-events-online/

[124] Lexico.com Online Dictionary, Dictionary.com and Oxford University Press. (2022). *Definition for: Belongingness*. Accessed: Mar. 31, 2022. [Online]. Available: https://www.lexico.com/definition/belongingness

[125] R. M. Lee and S. B. Robbins, ''Measuring belongingness: The social connectedness and the social assurance scales,'' *J. Counseling Psychol.*, vol. 42, no. 2, pp. 232–241, Apr. 1995, doi: 10.1037/0022-0167.42.2.232.

[126] G. P. Malone, D. R. Pillow, and A. Osman, ''The general belongingness scale (GBS): Assessing achieved belongingness,'' *Personality Individual Differences*, vol. 52, no. 3, pp. 311–316, Feb. 2012.

[127] F. Reer, W. Y. Tang, and T. Quandt, ''Psychosocial well-being and social media engagement: The mediating roles of social comparison orientation and fear of missing out,'' *New Media Soc.*, vol. 21, no. 7, pp. 1486–1505, Jul. 2019, doi: 10.1177/1461444818823719.

[128] K. Kırcaburun, C. M. Kokkinos, Z. Demetrovics, O. Király, M. D. Griffiths, and T. S. Çolak, ''Problematic online behaviors among adolescents and emerging adults: Associations between cyberbullying perpetration, problematic social media use, and psychosocial factors,'' *Int. J. Mental Health Addiction*, vol. 17, no. 4, pp. 891–908, Aug. 2019, doi: 10.1007/s11469-018-9894-8.

[129] R. Yavich, N. Davidovitch, and Z. Frenkel, ''Social media and loneliness–forever connected?'' *Higher Educ. Stud.*, vol. 9, no. 2, pp. 10–21, 2019.

[130] D. F. Sacco and M. M. Ismail, ''Social belongingness satisfaction as a function of interaction medium: Face-to-face interactions facilitate greater social belonging and interaction enjoyment compared to instant messaging,'' *Comput. Hum. Behav.*, vol. 36, pp. 359–364, Jul. 2014.

[131] A. A. Bennett, E. D. Campion, K. R. Keeler, and S. K. Keener, ''Videoconference fatigue? Exploring changes in fatigue after videoconference meetings during COVID-19,'' *J. Appl. Psychol.*, vol. 106, no. 3, p. 330, 2021.

[132] J. N. Bailenson, ''Nonverbal overload: A theoretical argument for the causes of zoom fatigue,'' *Technol., Mind, Behav.*, vol. 2, no. 1, pp. 1–6, Feb. 2021, doi: 10.1037/tmb0000030.

[133] N. Döring, K. D. Moor, M. Fiedler, K. Schoenenberg, and A. Raake, ''Videoconference fatigue: A conceptual analysis,'' *Int. J. Environ. Res. Public Health*, vol. 19, no. 4, p. 2061, Feb. 2022. [Online]. Available: https://www.mdpi.com/1660-4601/19/4/2061, doi: 10.3390/ijerph19042061.

[134] J. P. Kluck, F. Stoyanova, and N. C. Krämer, ''Putting the social back into physical distancing: The role of digital connections in a pandemic crisis,'' *Int. J. Psychol.*, vol. 56, no. 4, pp. 594–606, Aug. 2021.

[135] M. Marinucci, L. Pancani, N. Aureli, and P. Riva, ''Online social connections as surrogates of face-to-face interactions: A longitudinal study under COVID-19 isolation,'' *Comput. Hum. Behav.*, vol. 128, Mar. 2022, Art. no. 107102. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0747563221004258, doi: 10.1016/j.chb.2021.107102.

[136] Oxford Online Dictionary, Oxford University Press. (2020). *Definition for: Copresence*. Accessed: Mar. 31, 2022. [Online]. Available: https://www.oxfordreference.com/view/10.1093/oi/authority.20110803095638654

[137] S. Zhao, ''Toward a taxonomy of copresence,'' *Presence, Teleoperators Virtual Environ.*, vol. 12, no. 5, pp. 445–455, Oct. 2003, doi: 10.1162/105474603322761261.

[138] F. Biocca and B. Delaney, ''Immersive virtual reality technology,'' in *Communication in the Age of Virtual Reality*. Routledge, 1995, pp. 57–124.

[139] A. Perkis *et al.*, ''QUALINET white paper on definitions of immersive media experience (IMEx),'' May 2020, *arXiv:2007.07032*.

[140] F. Biocca, C. Harms, and J. K. Burgoon, ''Toward a more robust theory and measure of social presence: Review and suggested criteria,'' *Presence, Teleoperators Virtual Environments*, vol. 12, no. 5, pp. 456–480, Oct. 2003.

[141] R. Skarbez, F. P. Brooks, Jr., and M. C. Whitton, ''A survey of presence and related concepts,'' *ACM Comput. Surveys*, vol. 50, no. 6, pp. 1–39, Nov. 2018, doi: 10.1145/3134301.

[142] P. Rogers and M. Lea, ''Social presence in distributed group environments: The role of social identity,'' *Behav. Inf. Technol.*, vol. 24, no. 2, pp. 151–158, Mar. 2005.

[143] S. T. Bulu, ''Place presence, social presence, co-presence, and satisfaction in virtual worlds,'' *Comput. Educ.*, vol. 58, no. 1, pp. 154–161, Jan. 2012, doi: 10.1016/j.compedu.2011.08.024.

[144] A. Suh and J. Prophet, ''The state of immersive technology research: A literature analysis,'' *Comput. Hum. Behav.*, vol. 86, pp. 77–90, Sep. 2018.

[145] H. J. Smith and M. Neff, ''Communication behavior in embodied virtual reality,'' in *Proc. CHI Conf. Hum. Factors Comput. Syst. (CHI)*. New York, NY, USA: Association for Computing Machinery, Apr. 2018, pp. 1–2, doi: 10.1145/3173574.3173863.

[146] L. Morvan, A. Ovanessoff, M. C. Billiard, and F. Hintermann. (2014). *A Responsible Future for Immersive Technologies*. Accessed: Mar. 31, 2022. [Online]. Available: https://www.accenture.com/us-en/insights/technology/responsible-immersive-technologies

[147] C. Timmerer, M. Waltl, B. Rainer, and N. Murray, ''Sensory experience: Quality of experience beyond audio-visual,'' in *Quality of Experience*. Cham, Switzerland: Springer, 2014, pp. 351–365.

[148] A. Hamam, N. D. Georganas, and A. El Saddik, ''Effect of haptics on the quality of experience,'' in *Proc. IEEE Int. Symp. Haptic Audio Vis. Environ. Games*, Oct. 2010, pp. 1–6.

[149] R. Chaudhari, E. Altinsoy, and E. Steinbach, ''Haptics,'' in *Quality of Experience*. Cham, Switzerland: Springer, 2014, pp. 261–276.

[150] E. Stern, ''Individual differences in the learning potential of human beings,'' *npj Sci. Learn.*, vol. 2, no. 1, pp. 1–7, Dec. 2017.

[151] I. Wechsung, M. Schulz, K.-P. Engelbrecht, J. Niemann, and S. Möller, ''All users are (not) equal—The influence of user characteristics on perceived quality, modality choice and performance,'' in *Proc. Paralinguistic Inf. Its Integr. Spoken Dialogue Syst. Workshop*, R. L.-C. Delgado and T. Kobayashi, Eds. New York, NY, USA: Springer, 2011, pp. 175–186.

[152] G. Ghinea and S. Y. Chen, ''The impact of cognitive styles on perceptual distributed multimedia quality,'' *Brit. J. Educ. Technol.*, vol. 34, no. 4, pp. 393–406, Sep. 2003, doi: 10.1111/1467-8535.00337.

[153] B. Rainer, M. Waltl, E. Cheng, M. Shujau, C. Timmerer, S. Davis, I. Burnett, C. Ritz, and H. Hellwagner, ''Investigating the impact of sensory effects on the quality of experience and emotional response in web videos,'' in *Proc. 4th Int. Workshop Quality Multimedia Exper.*, Jul. 2012, pp. 278–283.

[154] A. Bhattacharya, W. Wu, and Z. Yang, ''Quality of experience evaluation of voice communication systems using affect-based approach,'' in *Proc. 19th ACM Int. Conf. Multimedia*. Scottsdale, AZ, USA: Association for Computing Machinery, 2011, pp. 929–932, doi: 10.1145/2072298.2071905.

[155] P. A. Andersen and L. K. Guerrero, Eds., *Handbook of Communication and Emotion: Research, Theory, Applications, and Contexts*. San Diego, CA, USA: Academic, 1997.

[156] J. A. Russell, ''A circumplex model of affect,'' *J. Personality Social Psychol.*, vol. 39, no. 6, pp. 1161–1178, Dec. 1980.

[157] N. A. Remington, L. R. Fabrigar, and P. S. Visser, ''Reexamining the circumplex model of affect,'' *J. Personality Social Psychol.*, vol. 79, no. 2, p. 286, 2000.

[158] J. Posner, J. A. Russell, and B. S. Peterson, ''The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology,'' *Develop. Psychopathol.*, vol. 17, no. 3, p. 715, Sep. 2005.

[159] D. Derks, A. H. Fischer, and A. E. R. Bos, ''The role of emotion in computer-mediated communication: A review,'' *Comput. Hum. Behav.*, vol. 24, no. 3, pp. 766–785, May 2008, doi: 10.1016/j.chb.2007.04.004.

[160] A. Batliner, B. Schuller, D. Seppi, S. Steidl, L. Devillers, L. Vidrascu, T. Vogt, V. Aharonson, and N. Amir, ''The automatic recognition of emotions in speech,'' in *Emotion-Oriented Systems: The Humaine Handbook*, R. Cowie, C. Pelachaud, and P. Petta, Eds. Berlin, Germany: Springer, 2011, pp. 71–99, doi: 10.1007/978-3-642-15184-2_6.

[161] J. Lassalle, L. Gros, and G. Coppin, ''Combination of physiological and subjective measures to assess quality of experience for audiovisual technologies,'' in *Proc. 3rd Int. Workshop Quality Multimedia Exper.*, Sep. 2011, pp. 13–18.

[162] K. Mitra, A. Zaslavsky, and C. Åhlund, ''A probabilistic context-aware approach for quality of experience measurement in pervasive systems,'' in *Proc. ACM Symp. Appl. Comput.* New York, NY, USA: Association for Computing Machinery, 2011, pp. 419–424, doi: 10.1145/1982185.1982276.

[163] L. He, M. Lech, N. C. Maddage, and N. B. Allen, ''Study of empirical mode decomposition and spectral analysis for stress and emotion classification in natural speech,'' *Biomed. Signal Process. Control*, vol. 6, no. 2, pp. 139–146, Apr. 2011, doi: 10.1016/j.bspc.2010.11.001.

[164] D.-H. Shin, ''Conceptualizing and measuring quality of experience of the Internet of Things: Exploring how quality is perceived by users,'' *Inf. Manage.*, vol. 54, no. 8, pp. 998–1011, 2017, doi: 10.1016/j.im.2017.02.006.

[165] S.-H. Liu, H.-L. Liao, and J. A. Pratt, ''Impact of media richness and flow on e-learning technology acceptance,'' *Comput. Educ.*, vol. 52, no. 3, pp. 599–607, Apr. 2009, doi: 10.1016/j.compedu.2008.11.002.

[166] S. Jumisko-Pyykkö and K. Väänänen-Vainio-Mattila, ''The role of audiovisual quality in mobile television,'' in *Proc. 2nd Int. Workshop Video Process. Qual. Metrics Consum. Electron.*, 2006, pp. 1–5.

[167] S. J. Breckler, ''Empirical validation of affect, behavior, and cognition as distinct components of attitude,'' *J. Personality Social Psychol.*, vol. 47, no. 6, p. 1191, 1984, doi: 10.1037/0022-3514.47.6.1191.

[168] U. Schiefele, "Topic interest, text representation, and quality of experience," *Contemp. Educ. Psychol.*, vol. 21, no. 1, pp. 3–18, Jan. 1996, doi: 10.1006/ceps.1996.0002.

[169] M. Hassenzahl, "User experience (UX): Towards an experiential perspective on product quality," in *Proc. 20th Int. Conf. Assoc. Francophone d'Interaction Homme-Mach. (IHM)*, 2008, pp. 11–15.

[170] M. Rokeach, *The Nature of Human Values*. New York, NY, USA: Free Press, 1973.

[171] A. J. Aubrey, D. Marshall, P. L. Rosin, J. Vendeventer, D. W. Cunningham, and C. Wallraven, "Cardiff conversation database (CCDb): A database of natural dyadic conversations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Portland, OR, USA, Jun. 2013, pp. 277–282, doi: 10.1109/CVPRW.2013.48.

[172] B. O'Conaill, S. Whittaker, and S. Wilbur, "Conversations over video conferences: An evaluation of the spoken aspects of video-mediated communication," *Hum.-Comput. Interact.*, vol. 8, no. 4, pp. 389–428, Dec. 1993.

[173] M. L. Knapp and J. A. Hall, *Nonverbal Communication in Human Interaction*, 7th ed. Boston, MA, USA: Cengage Learning, 2010.

[174] *Effect of Delays on Telemeeting Quality*, document ITU-T Rec. P.1305, International Standard, International Telecommunication Union, Geneva, Switzerland, 2016.

[175] B. Witmer and M. Singer, "Measuring presence in virtual environments: A presence questionnaire," *Presence, Teleoperators Virtual Environ.*, vol. 7, no. 3, pp. 225–240, Jun. 1998.

[176] S. Agrawal, S. Bech, K. Bærentsen, K. De Moor, and S. Forchhammer, "Method for subjective assessment of immersion in audiovisual experiences," *J. Audio Eng. Soc.*, vol. 69, no. 9, pp. 656–671, Sep. 2021.

[177] E. Nystad and A. Sebok, "A comparison of two presence measures based on experimental results," in *Proc. Presence Conf.*, Valencia, Spain, 2004, pp. 266–273.

[178] C. Youngblut and O. Huie, "The relationship between presence and performance in virtual environments: Results of a VERTS study," in *Proc. IEEE Virtual Reality*, Mar. 2003, pp. 277–278.

[179] A. Singla, S. Fremerey, W. Robitza, and A. Raake, "Measuring and comparing QoE and simulator sickness of omnidirectional videos in different head mounted displays," in *Proc. 9th Int. Conf. Quality Multimedia Exper. (QoMEX)*, May 2017, pp. 1–6.

[180] A. Singla, W. Robitza, and A. Raake, "Comparison of subjective quality evaluation methods for omnidirectional videos with DSIS and modified ACR," *Electron. Imag., Hum. Vis. Electron. Imag.*, vol. 2018, no. 14, pp. 1–6, 2018.

[181] A. Singla, S. Göring, A. Raake, B. Meixner, R. Koenen, and T. Buchholz, "Subjective quality evaluation of tile-based streaming for omnidirectional videos," in *Proc. 10th ACM Multimedia Syst. Conf.*, Jun. 2019, pp. 232–242.

[182] H. T. T. Tran, N. P. Ngoc, C. T. Pham, Y. J. Jung, and T. C. Thang, "A subjective study on QoE of 360 video for VR communication," in *Proc. IEEE 19th Int. Workshop Multimedia Signal Process. (MMSP)*, Oct. 2017, pp. 1–6.

[183] R. S. Kennedy, N. E. Lane, K. S. Berbaum, and M. G. Lilienthal, "Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness," *Int. J. Aviation Psychol.*, vol. 3, no. 3, pp. 203–220, 1993.

[184] N. Dużmańska, P. Strojny, and A. Strojny, "Can simulator sickness be avoided? A review on temporal aspects of simulator sickness," *Frontiers Psychol.*, vol. 9, p. 2132, Nov. 2018, doi: 10.3389/fpsyg.2018.02132.

[185] K. Hoeldtke and A. Raake, "Conversation analysis of multi-party conferencing and its relation to perceived quality," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Kyoto, Japan, Jun. 2011, pp. 1–5.

[186] M. Schmitt, S. Gunkel, P. Cesar, and D. Bulterman, "Asymmetric delay in video-mediated group discussions," in *Proc. 6th Int. Workshop Qual. Multimedia Exper. (QoMEX)*, Sep. 2014, pp. 19–24.

[187] J. Skowronek, F. Schiffner, and A. Raake, "On the influence of involvement on the quality of multiparty conferencing," in *Proc. 4th Int. Workshop Perceptual Quality Syst. (PQS)*. Vienna, Austria: International Speech Communication Association, Sep. 2013, pp. 141–146.

[188] M. C. B. Alastuey, "Synchronous-voice computer-mediated communication: Effects on pronunciation," *CALICO J.*, vol. 28, no. 1, pp. 1–20, Sep. 2010. [Online]. Available: https://www.jstor.org/stable/calicojournal.28.1.1

[189] S. Gahl, Y. Yao, and K. Johnson, "Why reduce? Phonological neighborhood density and phonetic reduction in spontaneous speech," *J. Memory Lang.*, vol. 66, no. 4, pp. 789–806, May 2012, doi: 10.1016/j.jml.2011.11.006.

[190] A. Loukina, M. Lopez, K. Evanini, D. Suendermann-Oeft, A. V. Ivanov, and K. Zechner, "Pronunciation accuracy and intelligibility of non-native speech," in *Proc. 16th Annu. Conf. Int. Speech Commun. Assoc.*, 2015, pp. 1917–1921.

[191] D. R. McCloy, R. A. Wright, and P. E. Souza, "Talker versus dialect effects on speech intelligibility: A symmetrical study," *Lang. Speech*, vol. 58, no. 3, pp. 371–386, Sep. 2015, doi: 10.1177/0023830914559234.

[192] D. S. Berry and J. S. Hansen, "Personality, nonverbal behavior, and interaction quality in female dyads," *Personality Social Psychol. Bull.*, vol. 26, no. 3, pp. 278–292, Mar. 2000, doi: 10.1177/0146167200265002.

[193] J. K. Burgoon, V. Manusov, and L. K. Guerrero, *Nonverbal Communication*, 2nd ed. Evanston, IL, USA: Routledge, 2021, doi: 10.4324/9781003095552.

[194] J. K. Burgoon, L. A. Stern, and L. Dillman, *Interpersonal Adaptation: Dyadic Interaction Patterns*. Cambridge, U.K.: Cambridge Univ. Press, 1995, doi: 10.1017/CBO9780511720314.

[195] K. Schoenenberg, A. Raake, and J. Skowronek, "A conversation analytic approach to the prediction of leadership in two to six-party audio conferences," in *Proc. 3rd Int. Workshop Quality Multimedia Exper.*, Sep. 2011, pp. 119–124.

[196] N. M. Suki, "Correlations of perceived flow, perceived system quality, perceived information quality, and perceived user trust on mobile social networking service (SNS) users' loyalty," *J. Inf. Technol. Res.*, vol. 5, no. 2, pp. 1–14, Apr. 2012, doi: 10.4018/jitr.2012040101.

[197] K. Schoenenberg, A. Raake, and J. Koeppe, "Why are you so slow?—Misattribution of transmission delay to attributes of the conversation partner at the far-end," *Int. J. Hum.-Comput. Stud.*, vol. 72, no. 5, pp. 477–487, May 2014, doi: 10.1016/j.ijhcs.2014.02.004.

[198] J. Segal, M. Smith, G. Boose, and J. Jaffe. (2020). Nonverbal communication. Help Guide, Santa Monica, CA, USA. Accessed: Mar. 31, 2022. [Online]. Available: https://www.helpguide.org/articles/relationships-communication/nonverbal-communication.htm

[199] A. Raake, K. Schoenenberg, J. Skowronek, and S. Egger, "Predicting speech quality based on interactivity and delay," in *Proc. 14th Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*. Lyon, France: International Speech Communication Association, Aug. 2013, pp. 1384–1388.

[200] M. J. Scott, S. C. Guntuku, Y. Huan, W. Lin, and G. Ghinea, "Modelling human factors in perceptual multimedia quality: On the role of personality and culture," in *Proc. 23rd ACM Int. Conf. Multimedia*. Brisbane, QLD, Australia: Association for Computing Machinery, 2015, pp. 481–490., doi: 10.1145/2733373.2806254.

[201] A. Sapru and H. Bourlard, "Automatic social role recognition in professional meetings using conditional random fields," in *Proc. 14th Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*. Lyon, France: International Speech Communication Association, Aug. 2013, pp. 2604–2608.

[202] M. Tomprou, Y. J. Kim, P. Chikersal, A. W. Woolley, and L. A. Dabbish, "Speaking out of turn: How video conferencing reduces vocal synchrony and collective intelligence," *PLoS ONE*, vol. 16, no. 3, Mar. 2021, Art. no. e0247655, doi: 10.1371//journal.pone.0247655.

[203] N. Kock, "Media richness or media naturalness? The evolution of our biological communication apparatus and its influence on our behavior toward E-communication tools," *IEEE Trans. Prof. Commun.*, vol. 48, no. 2, pp. 117–130, Jun. 2005.

[204] I. Zigurs and B. K. Buckland, "A theory of task/technology fit and group support systems effectiveness," *MIS Quart.*, vol. 22, no. 3, pp. 313–334, Sep. 1998.

[205] *Transmission Characteristics and Speech Quality Parameters of Hands-Free Terminals*, document ITU-T Rec. P.340, International Standard, International Telecommunication Union, Geneva, Switzerland, 2000.

[206] *IEEE Standard for Camera Phone Image Quality*, IEEE Standard 1858-2016 (Incorporating IEEE Std 1858-2016/Cor 1-2017, IEEE Standards Association Board of Governors, 2017, pp. 1–146, doi: 10.1109/IEEESTD.2017.7921676.

[207] L. Yao, A. DeVincenzi, A. Pereira, and H. Ishii, "FocalSpace: Multimodal activity tracking, synthetic blur and adaptive presentation for video conferencing," in *Proc. 1st Symp. Spatial User Interact.* Los Angeles, CA, USA: Association for Computing Machinery, 2013, p. 73–76, doi: 10.1145/2491367.2491377.

[208] A. B. Jeffery, J. D. Maes, and M. F. Bratton-Jeffery, "Improving team decision-making performance with collaborative modeling," *Team Perform. Management, Int. J.*, vol. 11, no. 1/2, pp. 40–50, Jan. 2005.

[209] O. DALY-JONES, A. Monk, and L. Watts, "Some advantages of video conferencing over high-quality audio conferencing: Fluency and awareness of attentional focus," *Int. J. Hum.-Comput. Stud.*, vol. 49, no. 1, pp. 21–58, Jul. 1998.

[210] S. Dubois, M. Boutin, and D. Sankoff, "The quantitative analysis of turntaking in multiparticipant conversations," *Univ. Pennsylvania Work. Papers Linguistics*, vol. 3, no. 1, 1996, Art. no. 20. [Online]. Available: https://repository.upenn.edu/pwpl/vol3/iss1/20

[211] E. A. Schegloff, "The relevance of repair to syntax-for-conversation," in *Discourse Syntax*. Leiden, The Netherlands: Brill, 1979, pp. 261–286.

[212] K. Brunnström, E. Dima, M. Andersson, M. Sjöström, T. Qureshi, and M. Johanson, "Quality of experience of hand controller latency in a virtual reality simulator," *Electron. Imag.*, vol. 2019, no. 12, pp. 218-1–218-9, 2019, doi: 10.2352/ISSN.2470-1173.2019.12.HVEI-218.

[213] F. Hammer, S. Egger-Lampl, and S. Möller, "Quality-of-user-experience: A position paper," *Qual. User Exper.*, vol. 3, no. 1, p. 9, 2018.

[214] T. Walton and M. Evans, "The role of human influence factors on overall listening experience," *Qual. User Exper.*, vol. 3, no. 1, p. 1, 2018.

[215] M. R. Quintero and A. Raake, "Is taking into account the subjects degree of knowledge and expertise enough when rating quality?" in *Proc. 4th Int. Workshop Qual. Multimedia Exper.*, Jul. 2012, pp. 194–199.

[216] F. Speranza, F. Poulin, R. Renaud, M. Caron, and J. Dupras, "Objective and subjective quality assessment with expert and non-expert viewers," in *Proc. 2nd Int. Workshop Qual. Multimedia Exper. (QoMEX)*, Jun. 2010, pp. 46–51.

[217] F. Rumsey, S. Zielinski, R. Kassier, and S. Bech, "Relationships between experienced listener ratings of multichannel audio quality and naïve listener preferences," *J. Acoust. Soc. Amer.*, vol. 117, no. 6, pp. 3832–3840, Jun. 2005.

[218] S. Bech and N. Zacharov, *Perceptual Audio Evaluation: Theory, Method and Application*. Chichester, U.K.: Wiley, 2006, doi: 10.1002/9780470869253.

[219] B. Lewcio, *Management of Speech and Video Telephony Quality in Heterogeneous Wireless Networks*. Cham, Switzerland: Springer, 2014.

[220] H. Santoso, M. Schrepp, A. Hinderks, and J. Thomaschewski, "Cultural differences in the perception of user experience," in *Mensch und Computer 2017—Tagungsband*, M. Burghardt, R. Wimmer, C. Wolff, and C. Womser-Hacker, Eds. Regensburg, Germany: Gesellschaft für Informatik e.V., 2017, pp. 267–272, doi: 10.18420/muc2017-mci-0272.

[221] R. E. Porter and L. A. Samovar, "Cultural influences on emotional expression: Implications for intercultural communication," in *Handbook of Communication and Emotion*, P. A. Andersen and L. K. Guerrero, Eds. San Diego, CA, USA: Academic, 1996, ch. 17, pp. 451–472, doi: 10.1016/B978-012057770-5/50019-9.

[222] T. H. Falk, C. Zheng, and W.-Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 7, pp. 1766–1774, Sep. 2010.

[223] Y. Zhu, I. Heynderickx, and J. A. Redi, "Understanding the role of social context and user factors in video quality of experience," *Comput. Hum. Behav.*, vol. 49, pp. 412–426, Aug. 2015, doi: 10.1016/j.chb.2015.02.054.

[224] *The Present State of Ultra-High Definition Television*, document ITU-R Rep. BT.2246-7, International Standardization Report, International Telecommunication Union, Geneva, Switzerland, 2020.

[225] *Video Quality Assessment of Streaming Services Over Reliable Transport for Resolutions up to 4K*, document ITU-T Rec. P.1204, International Standard, International Telecommunication Union, Geneva, Switzerland, 2020.

[226] H. Knoche and M. A. Sasse, "The big picture on small screens delivering acceptable video quality in mobile TV," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 5, no. 3, pp. 1–27, Aug. 2009.

[227] D. A. Norman, "Some observations on mental models," in *Mental Models*, D. Genter and A. L. Stevens, Eds. Hove, U.K.: Psychology Press, 1983, pp. 7–14.

[228] N. M. Yusoff and S. S. Salim, "Social-based versus shared situation awareness-based approaches to the understanding of team cognitive research in HCI," in *Proc. 3rd Int. Conf. User Sci. Eng. (i-USEr)*, Sep. 2014, pp. 281–286.

[229] G. Doherty-Sneddon, C. O'Malley, S. Garrod, A. Anderson, and E. Al, "Face-to-face and video-mediated communication: A comparison of dialogue structure and task performance," *J. Exp. Psychol., Appl.*, vol. 3, no. 2, pp. 105–125, 1997.

[230] J. Carletta, A. Isard, J. C. Kowtko, and G. Doherty-Sneddon, "Hcrc dialogue structure coding manual," Univ. Edinburgh, Edinburgh, U.K., Tech. Rep., 1996. [Online]. Available: http://www.lancaster.ac.uk./fass/projects/eagles/maptask.htm

[231] J. Carletta, A. Isard, S. Isard, G. Doherty-Sneddon, J. C. Kowtko, and A. H. Anderson, "The reliability of a dialogue structure coding scheme," *Comput. Linguistics*, vol. 23, no. 1, pp. 13–31, 1997.

[232] E. S. Veinott, J. Olson, G. M. Olson, and X. Fu, "Video helps remote work: Speakers who need to negotiate common ground benefit from seeing each other," in *Proc. CHI*, 1999, pp. 302–309.

[233] R. E. Kraut, S. R. Fussell, and J. Siegel, "Visual information as a conversational resource in collaborative physical tasks," *Hum.-Comput. Interact.*, vol. 18, nos. 1–2, pp. 13–49, Jun. 2003.

[234] H. H. Clark and S. E. Brennan, "Grounding in communication," in *Perspectives on Socially Shared Cognition*, L. B. Resnick, J. M. Levine, and S. D. Teasley, Eds. Washington, DC, USA: American Psychological Association, 1991, pp. 127–149.

[235] G. O'Neal, "Segmental repair and interactional intelligibility: The relationship between consonant deletion, consonant insertion, and pronunciation intelligibility in English as a lingua Franca in Japan," *J. Pragmatics*, vol. 85, pp. 122–134, Aug. 2015, doi: 10.1016/j.pragma.2015.06.013.

[236] C. Vaughn and A. Whitty, "Investigating the relationship between comprehensibility and social evaluation," *J. Second Lang. Pronunciation*, vol. 6, no. 3, pp. 483–504, Nov. 2020, doi: 10.1075/jslp.20022.vau.

[237] S. R. Fussell, "Social and cognitive processes in interpersonal communication: Implications for advanced telecommunications technologies," *Hum. Factors, J. Hum. Factors Ergonom. Soc.*, vol. 37, no. 2, pp. 228–250, Jun. 1995.

[238] J. Sidnell and T. Stivers, *The Handbook of Conversation Analysis*, vol. 121. Hoboken, NJ, USA: Wiley, 2012.

[239] E. M. Hoey and K. H. Kendrick, "Conversation analysis," in *Research Methods in Psycholinguistics and the Neurobiology of Language: A Practical Guide*, A. M. B. de Groot and P. Hagoort, Eds. Hoboken, NJ, USA: Wiley, 2017, pp. 151–173.

[240] P. T. Brady, "A technique for investigating on-off patterns of speech," *Bell Syst. Tech. J.*, vol. 44, no. 1, pp. 1–22, Jan. 1965.

[241] P. T. Brady, "A statistical analysis of on-off patterns in 16 conversations," *Bell Syst. Tech. J.*, vol. 47, no. 1, pp. 73–99, 1968.

[242] K. Otsuka, Y. Takemae, and J. Yamato, "A probabilistic inference of multiparty-conversation structure based on Markov-switching models of gaze patterns, head directions, and utterances," in *Proc. 7th Int. Conf. Multimodal Interfaces*. New York, NY, USA: Association for Computing Machinery, 2005, pp. 191–198, doi: 10.1145/1088463.1088497.

[243] P. T. Brady, "Effects of transmission delay on conversational behavior on echo-free telephone circuits," *Bell Syst. Tech. J.*, vol. 50, no. 1, pp. 115–134, Jan. 1971.

[244] F. Hammer, P. Reichl, and A. Raake, "Elements of interactivity in telephone conversations," in *Proc. 8th Int. Conf. Spoken Lang. Process. (INTERSPEECH)*, Jeju Island, South Korea, 2004, pp. 1741–1744.

[245] S. Egger, R. Schatz, K. Schoenenberg, A. Raake, and G. Kubin, "Same but different?—Using speech signal features for comparing conversational VoIP quality studies," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2012, pp. 1320–1324.

[246] K. Schoenenberg, "The quality of mediated-conversations under transmission delay," Ph.D. dissertation, Technische Univ. Berlin, Berlin, Germany, 2016.

[247] S. Möller, *Quality of Telephone-Based Spoken Dialogue Systems*. New York, NY, USA: Springer, 2005.

[248] W. Robitza, S. Schönfellner, and A. Raake, "A theoretical approach to the formation of quality of experience and user behavior in multimedia services," in *Proc. 5th ISCA/DEGA Workshop Perceptual Qual. Syst. (PQS)*, Aug. 2016, pp. 39–43.

[249] A. Silzle, "Quality taxonomies for auditory virtual environments," in *Proc. 122nd AES Conv.*, May 2007, pp. 1–18, Paper 6993.

[250] J. Skowronek and A. Raake, "Conceptual model of multiparty conferencing and telemeeting quality," in *Proc. 7th Int. Workshop Qual. Multimedia Exper. (QoMEX)*, Pilos, Greece, May 2015, pp. 1–6.

[251] J. Kim, D. Linsley, K. Thakkar, and T. Serre, "Disentangling neural mechanisms for perceptual grouping," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–12.

[252] A. Raake and H. Wierstorf, "Binaural evaluation of sound quality and quality of experience," in *The Technology of Binaural Understanding*. Cham, Switzerland: Springer, 2020, pp. 393–434.

[253] C. Spence and J. Driver, "Audiovisual links in exogenous covert spatial orienting," *Perception Psychophys.*, vol. 59, no. 1, pp. 1–22, Jan. 1997, doi: 10.3758/BF03206843.

[254] M. Kean and T. J. Crawford, "Cueing visual attention to spatial locations with auditory cues," *J. Eye Movement Res.*, vol. 2, no. 3, pp. 1–13, Dec. 2008, doi: 10.16910/jemr.2.3.4.

[255] O. Rummukainen and C. Mendonça, "Task-relevant spatialized auditory cues enhance attention orientation and peripheral target detection in natural scenes," *J. Eye Movement Res.*, vol. 9, no. 1, pp. 1–10, Jan. 2016, doi: 10.16910/jemr.9.1.4.

[256] E. Lombard, "Le signe de l'elevation de la voix," *Ann. Mal. de L'Oreille et du Larynx*, vol. 37, pp. 101–119, 1911. [Online]. Available: https://www.iqoe.org/library/16584

[257] H. Brumm and S. A. Zollinger, "The evolution of the lombard effect: 100 years of psychoacoustic research," *Behaviour*, vol. 148, nos. 11–13, pp. 1173–1198, 2011, doi: 10.1163/000579511X605759.

[258] A. W. Bronkhorst, "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions," *Acta Acustica*, vol. 86, no. 1, pp. 117–128, 2000.

[259] A. W. Bronkhorst, "The cocktail-party problem revisited: Early processing and selection of multi-talker speech," *Attention, Perception, Psychophys.*, vol. 77, no. 5, pp. 1465–1487, Jul. 2015, doi: 10.3758/s13414-015-0882-9.

[260] L. Itti and C. Koch, "Computational modelling of visual attention," *Nature Rev. Neurosci.*, vol. 2, no. 3, pp. 194–203, Mar. 2001.

[261] C. Kayser, C. I. Petkov, M. Lippert, and N. K. Logothetis, "Mechanisms for allocating auditory attention: An auditory saliency map," *Current Biol.*, vol. 15, no. 21, pp. 1943–1947, Nov. 2005.

[262] U. Engelke, H. Kaprykowsky, H. Zepernick, and P. Ndjiki-Nya, "Visual attention in quality assessment," *IEEE Signal Process. Mag.*, vol. 28, no. 6, pp. 50–59, Nov. 2011.

[263] J. Blauert and G. J. Brown, "Reflexive and reflective auditory feedback," in *The Technology of Binaural Understanding*, J. Blauert and J. Braasch, Eds. Cham, Switzerland: Springer, 2020, pp. 3–31, doi: 10.1007/978-3-030-00386-9_1.

[264] J. Piaget, *The Origins of Intelligence in Children*. New York, NY, USA: W. W. Norton & Company, 1952, doi: 10.1037/11494-000.

[265] A. Sackl, R. Schatz, and A. Raake, "More than I ever wanted or just good enough? User expectations and subjective quality perception in the context of networked multimedia services," *Qual. User Exper.*, vol. 2, no. 1, p. 3, 2017.

[266] A. Raake, M. Fiedler, K. Schoenenberg, K. De Moor, and N. Döring, "Technological factors influencing videoconferencing and zoom fatigue," 2022, *arXiv:2202.01740*.

[267] W. Schnotz and C. Kürschner, "A reconsideration of cognitive load theory," *Educ. Psychol. Rev.*, vol. 19, no. 4, pp. 469–508, Oct. 2007.

[268] R. Brunken, J. L. Plass, and D. Leutner, "Direct measurement of cognitive load in multimedia learning," *Educ. Psychol.*, vol. 38, no. 1, pp. 53–61, Jan. 2003.

[269] A.-F. N. M. Perrin, H. Xu, E. Kroupi, M. Řeřábek, and T. Ebrahimi, "Multimodal dataset for assessment of quality of experience in immersive multimedia," in *Proc. 23rd ACM Int. Conf. Multimedia*. Brisbane, QLD, Australia: Association for Computing Machinery, 2015, pp. 1007–1010, doi: 10.1145/2733373.2806387.

[270] A. Gaggioli, M. Bassi, and A. Fave, "Quality of experience in virtual environments," *Emerg. Commun.*, vol. 5, pp. 121–136, Sep. 2003.

[271] G. Riva, F. Mantovani, and A. Gaggioli, "Presence and rehabilitation: Toward second-generation virtual reality applications in neuropsychology," *J. Neuroeng. Rehabil.*, vol. 1, no. 1, p. 9, 2004.

[272] A. Gaggioli, "Quality of experience in real and virtual environments: Some suggestions for the development of positive technologies," in *Annual Review of Cybertherapy and Telemedicine 2012—Advanced Technologies in the Behavioral, Social and Neurosciences*, B. K. Wiederhold and G. Riva, Eds. Amsterdam, The Netherlands: IOS Press, 2012, pp. 177–181.

[273] K. Nowak, "Defining and differentiating copresence, social presence and presence as transportation," in *Proc. Presence Conf.*, Philadelphia, PA, USA, 2001, pp. 1–23.

[274] C. Neustaedter and S. Greenberg, "Intimacy in long-distance relationships over video chat," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.* Austin, TX, USA: Association for Computing Machinery, 2012, pp. 753–762, doi: 10.1145/2207676.2207785.

[275] F. De Simone, J. Li, H. G. Debarba, A. E. Ali, S. N. B. Gunkel, and P. Cesar, "Watching videos together in social virtual reality: An experimental study on user's QoE," in *Proc. IEEE Conf. Virtual Reality 3D User Interfaces (VR)*, Mar. 2019, pp. 890–891.

[276] A. Steed and R. Schroeder, "Collaboration in immersive and non-immersive virtual environments," in *Immersed in Media*. Cham, Switzerland: Springer, 2015, pp. 263–282.

[277] C. Youngblut, "Experience of presence in virtual environments," Inst. Defense Analyses, Alexandria, VA, USA, Tech. Rep. ADA427495, 2003. [Online]. Available: https://apps.dtic.mil/sti/citations/ADA427495

[278] S. Möller, M. Wältermann, and M.-N. Garcia, "Features of quality of experience," in *Quality of Experience—Advanced Concepts, Applications, Methods*, S. Möller and A. Raake, Eds. Cham, Switzerland: Springer, 2014, pp. 73–84.

[279] V.-V. Mattila, "Descriptive analysis and ideal point modelling of speech quality in mobile communications," in *Proc. 113th Audio Eng. Soc. (AES) Conv.*, Los Angeles, CA, USA, Oct. 2002, pp. 1–18, Paper 5704.

[280] *A Subjective Quality Test Methodology using Multiple Rating Scales*, document ITU-T Rec. P.806, International Standard, International Telecommunication Union, Geneva, Switzerland, 2014.

[281] F. Köster, *Multidimensional Analysis of Conversational Telephone Speech*. Cham, Switzerland: Springer, 2018.

[282] J. D. Carroll, "Individual differences and multidimensional scaling," in *Multidimensional Scaling: Theory and Applications in the Behavioral Sciences*, vol. 1, R. N. Shepard, A. K. Romney, and S. B. Nerlove, Eds. Seminar Press, 1972, pp. 55–105.

[283] H. Fastl and E. Zwicker, *Psychoacoustics: Facts and Models*. Berlin, Germany: Springer, 2007, doi: 10.1007/978-3-540-68888-4.

[284] B. C. Moore, *An Introduction to the Psychology of Hearing*, 7th ed. Leiden, The Netherlands: Brill, 2013.

[285] A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge, MA, USA: MIT Press, 1994.

[286] M. Hansen and B. Kollmeier, "Continuous assessment of time-varying speech quality," *J. Acoust. Soc. Amer.*, vol. 106, no. 5, pp. 2888–2899, Nov. 1999.

[287] L. Gros and N. Chateau, "Instantaneous and overall judgements for time-varying speech quality: Assessments and relationships," *Acta Acustica United With Acustica*, vol. 87, no. 3, pp. 367–377, 2001.

[288] B. Weiss, D. Guse, S. Möller, A. Raake, A. Borowiak, and U. Reiter, "Temporal development of quality of experience," in *Quality of Experience—Advanced Concepts, Applications, Methods*, S. Möller and A. Raake, Eds. Cham, Switzerland: Springer, 2014, pp. 133–147.

[289] *Parametric Bitstream-Based Quality Assessment of Progressive Download and Adaptive Audiovisual Streaming Services Over Reliable Transport—Quality Integration Module*, document ITU-T Rec. P.1203.3, International Standard, International Telecommunication Union, Geneva, Switzerland, 2019.

[290] W. Robitza, S. Göring, A. Raake, D. Lindegren, G. Heikkilä, J. Gustafsson, P. List, B. Feiten, U. Wüstenhagen, M.-N. Garcia, K. Yamagishi, and S. Broom, "HTTP adaptive streaming QoE estimation with ITU-T rec. P. 1203: Open databases and software," in *Proc. 9th ACM Multimedia Syst. Conf.*, 2018, pp. 466–471.

[291] N. Staelens, S. Moens, W. Van den Broeck, I. Marien, B. Vermeulen, P. Lambert, R. Van de Walle, and P. Demeester, "Assessing quality of experience of IPTV and video on demand services in real-life environments," *IEEE Trans. Broadcast.*, vol. 56, no. 4, pp. 458–466, Dec. 2010.

[292] M. Garau, D. Friedman, H. R. Widenfeld, A. Antley, A. Brogni, and M. Slater, "Temporal and spatial variations in presence: Qualitative analysis of interviews from an experiment on breaks in presence," *Presence, Teleoperators Virtual Environ.*, vol. 17, no. 3, pp. 293–309, Jun. 2008, doi: 10.1162/pres.17.3.293.

[293] *Subjective Evaluation of Conversational Quality*, document ITU-T Rec. P.805, International Standard, International Telecommunication Union, Geneva, Switzerland, 2007.

[294] *Subjective Performance Evaluation of Hands-Free Terminals*, document ITU-T Rec. P.832, International Standard, International Telecommunication Union, Geneva, Switzerland, 2000.

[295] *Subjective Test Methodology for Evaluating Speech Communication Systems That Include Noise Suppression Algorithm*, document ITU-T Rec. P.835, International Standard, International Telecommunication Union, Geneva, Switzerland, 2003.

[296] *Subjective Audiovisual Quality Assessment Methods for Multimedia Applications*, document ITU-T Rec. P.911, International Standard, International Telecommunication Union, Geneva, Switzerland, 1998.

[297] *Subjective Test Methodologies for 360° Video on Head-Mounted Displays*, document ITU-T Rec. P.919, International Standard, International Telecommunication Union, Geneva, Switzerland, 2020.

[298] *Interactive Test Methods for Audiovisual Communications*, document ITU-T Rec. P.920, International Standard, International Telecommunication Union, Geneva, Switzerland, 2000.

[299] *Interactive Test Methods for Audiovisual Communications*, document ITU-T Rec. P.920, International Standard, International Telecommunication Union, Geneva, Switzerland, 2000.

[300] *Methodologies for the Subjective Assessment of the Quality of Television Images*, document ITU-R Rec. BT.500, International Standardization Report, International Telecommunication Union, Geneva, Switzerland, 2019.

[301] *Method for Objective Measurements of Perceived Audio Quality*, document ITU-R Rec. BS.1387, International Standardization Report, International Telecommunication Union, Geneva, Switzerland, 2001.

[302] *Wideband Emodel*, document ITU-T Rec. G.107.1, International Standard, International Telecommunication Union, Geneva, Switzerland, 2019.

[303] *Fullband Emodel*, document ITU-T Rec. G.107.2, International Standard, International Telecommunication Union, Geneva, Switzerland, 2019.

[304] *Opinion Model for Video-Telephony Applications*, document ITU-T Rec. G.1070, International Standard, International Telecommunication Union, Geneva, Switzerland, 2018.

[305] *Objective Perceptual Video Quality Measurement Techniques for Digital Cable Television in the Presence of a Full Reference*, document ITU-T Rec. J.144, International Standard, International Telecommunication Union, Geneva, Switzerland, 2004.

[306] *Perceptual Visual Quality Measurement Techniques for Multimedia Services Over Digital Cable Television Networks in the Presence of a Reduced Bandwidth Reference*, document ITU-T Rec. J.246, International Standard, International Telecommunication Union, Geneva, Switzerland, 2008.

[307] *Objective Perceptual Multimedia Video Quality Measurement in the Presence of a Full Reference*, document ITU-T Rec. J.247, International Standard, International Telecommunication Union, Geneva, Switzerland, 2008.

[308] *Objective Perceptual Multimedia Video Quality Measurement of HDTV for Digital Cable Television in the Presence of a Full Reference*, document ITU-T Rec. J.341, International Standard, International Telecommunication Union, Geneva, Switzerland, 2016.

[309] *Analysis and Interpretation of INMD Voice-Service Measurements*, document ITU-T Rec. P.562, International Standard, International Telecommunication Union, Geneva, Switzerland, 2004.

[310] *Single-Ended Method for Objective Speech Quality Assessment in Narrow-Band Telephony Applications*, document ITU-T Rec. P.563, International Standard, International Telecommunication Union, Geneva, Switzerland, 2004.

[311] *Conformance Testing for Voice Over IP Transmission Quality Assessment Models*, document ITU-T Rec. P.564, International Standard, International Telecommunication Union, Geneva, Switzerland, 2007.

[312] *Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-End Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs*, document ITU-T Rec. P.862, International Standard, International Telecommunication Union, Geneva, Switzerland, 2001.

[313] *Wideband Extension to Recommendation P.862 for the Assessment of Wideband Telephone Networks and Speech Codecs*, document ITU-T Rec. P.862.2, International Standard, International Telecommunication Union, Geneva, Switzerland, 2007.

[314] *Perceptual Objective Listening Quality Prediction*, document ITU-T Rec. P.863, International Standard, International Telecommunication Union, Geneva, Switzerland, 2018.

[315] *Parametric Nonintrusive Assessment of Audiovisual Media Streaming Quality*, document ITU-T Rec. P.1201, International Standard, International Telecommunication Union, Geneva, Switzerland, 2012.

[316] *Parametric Non-Intrusive Bitstream Assessment of Video Media Streaming Quality*, document ITU-T Rec. P.1202, International Standard, International Telecommunication Union, Geneva, Switzerland, 2012.

[317] *Video Quality Assessment of Streaming Services Over Reliable Transport for Resolutions up to 4K With Access to Full Bitstream Information*, document ITU-T Rec. P.1204.3, International Standard, International Telecommunication Union, Geneva, Switzerland, 2020.

[318] *Video Quality Assessment of Streaming Services Over Reliable Transport for Resolutions up to 4K With Access to Full and Reduced Reference Pixel Information*, document ITU-T Rec. P.1204.4, International Standard, International Telecommunication Union, Geneva, Switzerland, 2020.

[319] *Video Quality Assessment of Streaming Services Over Reliable Transport for Resolutions up to 4K With Access to Transport and Received Pixel Information*, document ITU-T Rec. P.1204.5, International Standard, International Telecommunication Union, Geneva, Switzerland, 2020.

[320] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004, doi: 10.1109/TIP.2003.819861.

[321] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara, "Toward a practical perceptual video quality metric," *Netflix Technol. Blog*, vol. 62, no. 2, Jun. 2016. Accessed: Mar. 31, 2022. [Online]. Available: https://netflixtechblog.com/toward-a-practical-perceptual-video-quality-metric-653f208b9652

[322] Z. Li, C. Bampis, J. Novak, A. Aaron, K. Swanson, A. Moorthy, and J. D. Cock, "VMAF: The journey continues," *Netflix Technol. Blog*, vol. 25, Oct. 2018. Accessed: Mar. 31, 2022. [Online]. Available: https://netflixtechblog.com/vmaf-the-journey-continues-44b51ee9ed12

[323] C. G. Bampis, Z. Li, and A. C. Bovik, "Spatiotemporal feature integration and model fusion for full reference video quality assessment," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 8, pp. 2256–2270, Sep. 2019, doi: 10.1109/TCSVT.2018.2868262.

[324] G. Mittag and S. Möller, "Non-intrusive speech quality assessment for super-wideband speech communication networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 7125–7129, doi: 10.1109/ICASSP.2019.8683770.

[325] M. Chinen, F. S. C. Lim, J. Skoglund, N. Gureev, F. O'Gorman, and A. Hines, "ViSQOL v3: An open source production ready objective speech and audio metric," in *Proc. 12th Int. Conf. Quality Multimedia Exper. (QoMEX)*, May 2020, pp. 1–6, doi: 10.1109/QoMEX48832.2020.9123150.

[326] Z. Akhtar, K. Siddique, A. Rattani, S. L. Lutfi, and T. H. Falk, "Why is multimedia quality of experience assessment a challenging problem?" *IEEE Access*, vol. 7, pp. 117897–117915, 2019, doi: 10.1109/ACCESS.2019.2936470.

[327] *Subjective Evaluation of Speech Quality With a Crowdsourcing Approach*, document ITU-T Rec. P.808, International Standard, International Telecommunication Union, Geneva, Switzerland, 2018.

[328] *Mean Opinion Score (MOS) Terminology*, document ITU-T Rec. P.800.1, International Standard, International Telecommunication Union, Geneva, Switzerland, 2016.

[329] M. Adel, H. Assem, B. Jennings, D. Malone, J. Dunne, and P. O'Sullivan, "Improved E-model for monitoring quality of multi-party VoIP communications," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Atlanta, GA, USA, Dec. 2013, pp. 1180–1185, doi: 10.1109/GLOCOMW.2013.6825153.

[330] J. Skowronek and A. Raake, "On the quality perception of multiparty conferencing calls," in *Proc. 10th Int. Conf. Quality Multimedia Exper. (QoMEX)*, May 2018, pp. 1–6, doi: 10.1109/QoMEX.2018.8463378.

[331] *Reference Guide to Quality of Experience Assessment Methodologies*, document ITU-T Rec. G.1011, International Standard, International Telecommunication Union, Geneva, Switzerland, 2015.

[332] *Spatial Audio Meetings Quality Evaluation*, document ITU-T Rec. P.1310, International Standard, International Telecommunication Union, Geneva, Switzerland, 2017.

[333] International Telecommunication Union. (2020). *ITU-R Recommendations—BS-Series: Broadcasting Service (Sound)*. Accessed: Mar. 31, 2022. [Online]. Available: https://www.itu.int/rec/R-REC-BS/en

[334] International Telecommunication Union. (2020). *ITU-R Recommendations—BT-Series: Broadcasting Service (Television)*. Accessed: Mar. 31, 2022. [Online]. Available: https://www.itu.int/rec/R-REC-BT/en

[335] S. Möller and F. Köster, "Review of recent standardization activities in speech quality of experience," *Qual. User Exper.*, vol. 2, no. 1, p. 9, 2017, doi: 10.1007/s41233-017-0012-7.

[336] S. Möller and A. Raake, Eds., *Quality of Experience—Advanced Concepts, Applications, Methods*. Cham, Switzerland: Springer, 2014.

[337] N. Barman and M. G. Martini, "QoE modeling for HTTP adaptive video streaming—A survey and open challenges," *IEEE Access*, vol. 7, pp. 30831–30859, 2019.

[338] P. Juluri, V. Tamarapalli, and D. Medhi, "Measurement of quality of experience of video-on-demand services: A survey," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 1, pp. 401–418, 1st Quart., 2016.

[339] A. Raake, S. Borer, S. M. Satti, J. Gustafsson, R. R. R. Rao, S. Medagli, P. List, S. Göring, D. Lindero, W. Robitza, and G. Heikkilä, "Multi-model standard for bitstream-, pixel-based and hybrid video quality assessment of UHD/4K: ITU-T P. 1204," *IEEE Access*, vol. 8, pp. 193020–193049, 2020.

[340] *Method for Determining the Intelligibility of Multiple Concurrent Talkers*, document ITU-T Rec. P.1311, International Standard, International Telecommunication Union, Geneva, Switzerland, 2014.

[341] M. Orduna, P. Pérez, J. Gutiérrez, and N. García, "Methodology to assess quality, presence, empathy, attitude, and attention in 360-degree videos for immersive communications," *IEEE Trans. Affect. Comput.*, early access, Feb. 14, 2022, doi: 10.1109/TAFFC.2022.3149162.

[342] N. R. French and J. C. Steinberg, "Factors governing the intelligibility of speech sounds," *J. Acoust. Soc. Amer.*, vol. 19, no. 1, pp. 90–119, Jan. 1947.

[343] H. J. M. Steeneken and T. Houtgast, "A physical method for measuring speech-transmission quality," *J. Acoust. Soc. Amer.*, vol. 67, no. 1, pp. 318–326, Jan. 1980.

[344] *Methods for Calculation of the Speech Intelligibility Index*, Standard ANSI S3.5-1997 (R2007), American National Standards Institute, New York, NY, USA, 1997.

[345] J. B. Allen, *Articulation and Intelligibility*. San Rafael, CA, USA: Morgan & Claypool, 2006, doi: 10.2200/S00004ED1V01Y200508SAP001.

[346] F. Schiffner, J. Skowronek, and A. Raake, "On the impact of speech intelligibility on speech quality in the context of voice over IP telephony," in *Proc. 6th Int. Workshop Qual. Multimedia Exper. (QoMEX)*, Sep. 2014, pp. 59–60.

[347] J. E. Preminger and D. J. V. Tasell, "Quantifying the relation between speech quality and speech intelligibility," *J. Speech, Lang., Hearing Res.*, vol. 38, no. 3, pp. 714–725, Jun. 1995, doi: 10.1044/jshr.3803.714.

[348] P. C. Loizou and G. Kim, "Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 1, pp. 47–56, Jan. 2011.

[349] A. Warzybok, I. Kodrasi, J. O. Jungmann, E. Habets, T. Gerkmann, A. Mertins, S. Doclo, B. Kollmeier, and S. Goetze, "Subjective speech quality and speech intelligibility evaluation of single-channel dereverberation algorithms," in *Proc. 14th Int. Workshop Acoustic Signal Enhancement (IWAENC)*, Sep. 2014, pp. 332–336.

[350] M. M. Kabir, M. F. Mridha, J. Shin, I. Jahan, and A. Q. Ohi, "A survey of speaker recognition: Fundamental theories, recognition methods and opportunities," *IEEE Access*, vol. 9, pp. 79236–79263, 2021, doi: 10.1109/ACCESS.2021.3084299.

[351] R. Mohd Hanifa, K. Isa, and S. Mohamad, "A review on speaker recognition: Technology and challenges," *Comput. Electr. Eng.*, vol. 90, Mar. 2021, Art. no. 107005, doi: 10.1016/j.compeleceng.2021.107005.

[352] Z. Bai and X.-L. Zhang, "Speaker recognition based on deep learning: An overview," *Neural Netw.*, vol. 140, pp. 65–99, Aug. 2021, doi: 10.1016/j.neunet.2021.03.004.

[353] G. Chaudhary, S. Srivastava, and S. Bhardwaj, "Feature extraction methods for speaker recognition: A review," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 31, no. 12, Dec. 2017, Art. no. 1750041, doi: 10.1142/S0218001417500410.

[354] J. H. L. Hansen and T. Hasan, "Speaker recognition by machines and humans: A tutorial review," *IEEE Signal Process. Mag.*, vol. 32, no. 6, pp. 74–99, Nov. 2015, doi: 10.1109/MSP.2015.2462851.

[355] R. Jahangir, Y. W. Teh, H. F. Nweke, G. Mujtaba, M. A. Al-Garadi, and I. Ali, "Speaker identification through artificial intelligence techniques: A comprehensive review and research challenges," *Expert Syst. Appl.*, vol. 171, Jun. 2021, Art. no. 114591, doi: 10.1016/j.eswa.2021.114591.

[356] S. S. Tirumala, S. R. Shahamiri, A. S. Garhwal, and R. Wang, "Speaker identification features extraction methods: A systematic review," *Expert Syst. Appl.*, vol. 90, pp. 250–271, Dec. 2017, doi: 10.1016/j.eswa.2021.114591.

[357] L. F. Gallardo, *Human and Automatic Speaker Recognition over Telecommunication Channels* (T-Labs Series in Telecommunication Services). Singapore: Springer, 2016. [Online]. Available: https://link.springer.com/book/10.1007/978-981-287-727-7, doi: 10.1007/978-981-287-727-7.

[358] R. Huber, J. Ooster, and B. Meyer, "Single-ended speech quality prediction based on automatic speech recognition," *J. Audio Eng. Soc.*, vol. 66, no. 10, pp. 759–769, Oct. 2018, doi: 10.17743/jaes.2018.0041.

[359] G. Mittag, *Deep Learning Based Speech Quality Prediction*. Cham, Switzerland: Springer, 2012, doi: 10.1007/978-3-030-91479-0.

[360] T. Polzehl, S. Moller, and F. Metze, "Automatically assessing personality from speech," in *Proc. IEEE 4th Int. Conf. Semantic Comput.*, Sep. 2010, pp. 134–140, doi: 10.1109/ICSC.2010.41.

[361] T. Polzehl, *Personality in Speech—Assessment and Automatic Classification* (T-Labs Series in Telecommunication Services). Cham, Switzerland: Springer, 2014.

[362] L. H. Gilpin, D. M. Olson, and T. Alrashed, "Perception of speaker personality traits using speech signals," in *Proc. CHI EA*. Montreal QC, Canada: Association for Computing Machinery, 2018, p. 1–6.

[363] A. Guidi, C. Gentili, E. P. Scilingo, and N. Vanello, "Analysis of speech features and personality traits," *Biomed. Signal Process. Control*, vol. 51, pp. 1–7, May 2019, doi: 10.1016/j.bspc.2019.01.027.

[364] B. L. Smith, B. L. Brown, W. J. Strong, and A. C. Rencher, "Effects of speech rate on personality perception," *Lang. Speech*, vol. 18, no. 2, pp. 145–152, Apr. 1975.

[365] T. R. Schussler, A. Smithson, J. Smithson, A. Bonasio, M. Sage, F. Ozkan, K. Wyman, M. Davis, D. Woodruff, B. Erwin, A. Colgan, and L. Spring, "A global resource guide to XR collaboration," XR Ignite, Toronto, ON, Canada, en-US. Tech. Rep., 2020. Accessed: Mar. 31, 2022. [Online]. Available: http://xrcollaboration.com/

[366] M. Slater and M. V. Sanchez-Vives, "Enhancing our lives with immersive virtual reality," *Front. Robot. AI*, vol. 3, p. 74, Dec. 2016.

[367] J. McVeigh-Schultz and K. Isbister, "The case for 'weird social' in VR/XR: A vision of social superpowers beyond meatspace," in *Proc. Extended Abstr. CHI Conf. Hum. Factors Comput. Syst.*, May 2021, pp. 1–10.

[368] J. A. Lewnard and N. C. Lo, "Scientific and ethical basis for social-distancing interventions against COVID-19," *Lancet Infect. Dis.*, vol. 20, no. 6, pp. 631–633, 2020.

[369] E. Bin, C. Andruetto, Y. Susilo, and A. Pernestål, "The trade-off behaviours between virtual and physical activities during the first wave of the COVID-19 pandemic period," *Eur. Transp. Res. Rev.*, vol. 13, no. 1, pp. 1–19, Dec. 2021, doi: 10.1016/S1473-3099(20)30190-0.

[370] G. Gonçalves, M. Melo, J. Vasconcelos-Raposo, and M. Bessa, "Impact of different sensory stimuli on presence in credible virtual environments," *IEEE Trans. Vis. Comput. Graphics*, vol. 26, no. 11, pp. 3231–3240, Nov. 2020, doi: 10.1109/TVCG.2019.2926978.

[371] *5G; Extended Reality (XR) in 5G (3GPP TR 26.928 Version 16.0.0 Release 16)*, document ETSI TR 126 928 V16.0.0 (2020-11), International Standardization Report, European Telecommunications Standards Institute, Sophia Antipolis Cedex, France, 2020.

[372] T. D. Koninck. (2020). How social XR (extended reality) reduces distances. Nederlandse Organisatie voor Toegepast Natuurwetenschappelijk Onderzoek (TNO). Accessed: Mar. 31, 2022. [Online]. Available: https://www.tno.nl/en/focus-areas/information-communication-technology/roadmaps/fast-open-infrastructures/social-xr-extended-reality/

[373] T. D. Parsons, A. Gaggioli, and G. Riva, "Extended reality for the clinical, affective, and social neurosciences," *Brain Sci.*, vol. 10, no. 12, p. 922, Nov. 2020, doi: 10.3390/brainsci10120922.

[374] C. Fink, M. Billinghurst, J. Kotkin, B. Cloobeck, J. Colby, J. Dillard, D. Liberman, B. Fink, A. Liu, H. Morris, B. Sullivan, and T. Xu, *Remote Collaboration, Virtual Conferences, and the Future of Work*, 1st ed. Convergence Press, Jun. 2020. [Online]. Available: https://www.amazon.com/Remote-Collaboration-Virtual-Conferencing-Future-ebook/dp/B08D6V52D2/ref=sr_1_1?crid=1FC1ALZAUU33Y&keywords=Remote+Collaboration+%26+Virtual+Conferencing%3A+The+Future+Of+Work&qid=1654689780&sprefix=remote+collaboration+%26+virtual+conferencing+the+future+of+work%2Caps%2C185&sr=8-1

[375] J. Wiecha, R. Heyden, E. Sternthal, and M. Merialdi, "Learning in a virtual world: Experience with using second life for medical education," *J. Med. Internet Res.*, vol. 12, no. 1, p. e1, Jan. 2010.

[376] J. McVeigh-Schultz, A. Kolesnichenko, and K. Isbister, "Shaping pro-social interaction in VR: An emerging design framework," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, May 2019, pp. 1–12.

[377] S. Cho, S.-W. Kim, J. Lee, J. Ahn, and J. Han, "Effects of volumetric capture avatars on social presence in immersive virtual environments," in *Proc. IEEE Conf. Virtual Reality 3D User Interface (VR)*, Mar. 2020, pp. 26–34, doi: 10.1109/VR46266.2020.00-84.

[378] G. Gamelin, A. Chellali, S. Cheikh, A. Ricca, C. Dumas, and S. Otmane, "Point-cloud avatars to improve spatial communication in immersive collaborative virtual environments," *Pers. Ubiquitous Comput.*, vol. 25, no. 3, pp. 467–484, Jun. 2021, doi: 10.1007/s00779-020-01431-1.

[379] M. Kyan, G. Sun, H. Li, L. Zhong, P. Muneesawang, N. Dong, B. Elder, and L. Guan, "An approach to ballet dance training through MS Kinect and visualization in a CAVE virtual reality environment," *ACM Trans. Intell. Syst. Technol.*, vol. 6, no. 2, pp. 1–37, May 2015.

[380] M. Gorzynski, M. Derocher, and A. S. Mitchell, "The Halo B2B studio," in *Media Space 20 + Years of Mediated Life*, S. Harrison, Ed. London, U.K.: Springer, 2009, pp. 357–368, doi: 10.1007/978-1-84882-483-6_22.

[381] J. G. Apostolopoulos, P. A. Chou, B. Culbertson, T. Kalker, M. D. Trott, and S. Wee, "The road to immersive communication," *Proc. IEEE*, vol. 100, no. 4, pp. 974–990, Apr. 2012, doi: 10.1109/JPROC.2011.2182069.

[382] K. O'hara, J. Kjeldskov, and J. Paay, "Blended interaction spaces for distributed team collaboration," *ACM Trans. Comput.-Hum. Interact.*, vol. 18, no. 1, pp. 1–28, Apr. 2011, doi: 10.1145/1959022.1959025.

[383] S. Orts-Escolano *et al.*, "Holoportation: Virtual 3D teleportation in real-time," in *Proc. 29th Annu. Symp. User Interface Softw. Technol.*, Oct. 2016, pp. 741–754.

[384] L. Yoon, D. Yang, J. Kim, C. Chung, and S.-H. Lee, "Placement retargeting of virtual avatars to dissimilar indoor environments," *IEEE Trans. Vis. Comput. Graphics*, vol. 28, no. 3, pp. 1619–1633, Mar. 2022, doi: 10.1109/TVCG.2020.3018458.

[385] C. Ennis and C. O'Sullivan, "Perceptually plausible formations for virtual conversers," *Comput. Animation Virtual Worlds*, vol. 23, nos. 3–4, pp. 321–329, May 2012, doi: 10.1002/cav.1453.

[386] K. Zibrek and R. McDonnell, "Social presence and place illusion are affected by photorealism in embodied vr," in *Proc. Motion, Interact. Games*. New York, NY, USA: Association for Computing Machinery, 2019, Art. no. 13, doi: 10.1145/3359566.3360064.

[387] T. Zhan, K. Yin, J. Xiong, Z. He, and S.-T. Wu, "Augmented reality and virtual reality displays: Perspectives and challenges," *iScience*, vol. 23, no. 8, p. 101397, 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S258900422030585X, doi: 10.1016/j.isci.2020.101397.

[388] M. Wong and R. Duraiswami, "Shared-space: Spatial audio and video layouts for videoconferencing in a virtual room," in *Proc. Immersive 3D Audio, From Archit. Automot. (IDA)*, Sep. 2021, pp. 1–6, doi: 10.1109/I3DA48870.2021.9610974.

[389] W. P. de Bruijn and M. M. Boone, "Application of wave field synthesis in life-size videoconferencing," in *Proc. Audio Eng. Soc. Conv.*, 2003, pp. 1–17, Paper 5801.

[390] W. P. De Bruijn, "Application of wave field synthesis in videoconferencing," Ph.D. dissertation, Lab. Acoust. Imag. Sound Control, Fac. Appl. Sci., Delft Univ. Technol., Delft, The Netherlands, 2004.

[391] J. Hill. (2018). Bluejeans brings spatial audio calls to conference rooms. Blue Jeans Network. Accessed: Mar. 31, 2022. [Online]. Available: https://www.bluejeans.com/blog/bluejeans-spatial-audio-calls-to-conference-rooms

[392] British Telecom. (2020). *BT MEETME With Dolby Voice—User Guide*. Accessed: Mar. 31, 2022. [Online]. Available: https://www.globalservices.bt.com/en/my-account/support/collaboration/meetme-with-dolby-voice/user-guides/inviting-people-to-your-meeting

[393] *Cisco TelePresence TX9000 Series*, Cisco, San Jose, CA, USA, 2012.

[394] *Cisco IX5000 Series. Data Sheet*, Cisco, San Jose, CA, USA, 2018.

[395] S. Werner, F. Klein, T. Mayenfels, and K. Brandenburg, "A summary on acoustic room divergence and its effect on externalization of auditory events," in *Proc. 8th Int. Conf. Quality Multimedia Exper. (QoMEX)*, Jun. 2016, pp. 1–6.

[396] V. Best, R. Baumgartner, M. Lavandier, P. Majdak, and N. Kopčo, "Sound externalization: A review of recent research," *Trends Hearing*, vol. 24, pp. 1–14, Jan. 2020, doi: 10.1177/2331216520948390.

[397] S. Werner, F. Klein, and K. Brandenburg, "Influence of spatial complexity and room acoustic disparity on perception of quality features using a binaural synthesis system," in *Proc. 7th Int. Workshop Quality Multimedia Exper. (QoMEX)*, May 2015, pp. 1–6.

[398] F. Klein, S. Werner, and T. Mayenfels, "Influences of training on externalization of binaural synthesis in situations of room divergence," *J. Audio Eng. Soc.*, vol. 65, no. 3, pp. 178–187, Mar. 2017.

[399] B. Jones, Y. Zhang, P. N. Y. Wong, and S. Rintel, "Belonging there: VROOM-ing into the uncanny valley of XR telepresence," *Proc. ACM Hum.-Comput. Interact.*, vol. 5, Apr. 2021, Art. no. 59, doi: 10.1145/3449133.

[400] M. Mori, K. F. MacDorman, and N. Kageki, "The uncanny valley," *IEEE Robot. Autom. Mag.*, vol. 19, no. 2, pp. 98–100, Jun. 2012, doi: 10.1109/MRA.2012.2192811.

[401] Khronos Group. (2022). *Open XR*. Accessed: Mar. 31, 2022. [Online]. Available: https://www.khronos.org/openxr/

[402] D. Van Den Berg, R. Glans, D. De Koning, F. Kuipers, J. Lugtenburg, K. Polachan, P. Venkata, C. Singh, B. Turkovic, and B. Van Wijk, "Challenges in haptic communications over the tactile internet," *IEEE Access*, vol. 5, pp. 23502–23518, 2017, doi: 10.1109/ACCESS.2017.2764181.

[403] W. Wirth, T. Hartmann, S. Böcking, P. Vorderer, C. Klimmt, H. Schramm, T. Saari, J. Laarni, N. Ravaja, F. R. Gouveia, F. Biocca, A. Sacau, L. Jäncke, T. Baumgartner, and P. Jäncke, "A process model of the formation of spatial presence experiences," *Media Psychol.*, vol. 9, no. 3, pp. 493–525, May 2007, doi: 10.1080/15213260701283079.

[404] B. K. Wiederhold, D. P. Jang, M. Kaneda, I. Cabral, Y. Lurie, T. May, I. Y. Kim, M. D. Wiederhold, and S. I. Kim, "An investigation into physiological responses in virtual environments: An objective measurement of presence," in *Towards Cyberpsychology: Mind, Cognitions and Society in the Internet Age*, G. Riva and C. Galimberti, Eds. Amsterdam, The Netherlands: IOS Press, 2001, pp. 175–183.

[405] C. Keighrey, R. Flynn, S. Murray, and N. Murray, "A physiology-based QoE comparison of interactive augmented reality, virtual reality and tablet-based applications," *IEEE Trans. Multimedia*, vol. 23, pp. 333–341, 2021.

[406] R. A. Ruddle and S. Lessels, "Three levels of metric for evaluating wayfinding," *Presence: Teleoperators Virtual Environ.*, vol. 15, no. 6, pp. 637–654, Dec. 2006.

[407] O. Rummukainen, S. Schlecht, A. Plinge, and E. A. Habets, "Evaluating binaural reproduction systems from behavioral patterns in a virtual reality—A case study with impaired binaural cues and tracking latency," in *Proc. Audio Eng. Soc. Conv.* New York, NY, USA: Audio Engineering Society, 2017, pp. 1–8, Paper 9895.

[408] L. Aymerich-Franch and E. Fosch-Villaronga, "A self-guiding tool to conduct research with embodiment technologies responsibly," *Frontiers Robot. AI*, vol. 7, p. 22, Feb. 2020. [Online]. Available: https://www.frontiersin.org/article/10.3389/frobt.2020.00022, doi: 10.3389/frobt.2020.00022.

[409] J. Li, V. Vinayagamoorthy, R. Schwartz, W. IJsselsteijn, D. A. Shamma, and P. Cesar, "Social VR: A new medium for remote communication and collaboration," in *Proc. Extended Abstr. CHI Conf. Hum. Factors Comput. Syst.*, Apr. 2020, pp. 1–8, doi: 10.1145/3334480.3375160.

[410] P. B. Lowry, J. Zhang, C. Wang, and M. Siponen, "Why do adults engage in cyberbullying on social media? An integration of online disinhibition and deindividuation effects with the social structure and social learning model," *Inf. Syst. Res.*, vol. 27, no. 4, pp. 962–986, Dec. 2016, doi: 10.1287/isre.2016.0671.

[411] K. A. McCarthy, "Is rudeness really that common? An exploratory study of incivility at work," *J. Organizational Comput. Electron. Commerce*, vol. 26, no. 4, pp. 364–374, Oct. 2016.

[412] XR Association. (2020). *Research & Best Practices*. Accessed: Mar. 31, 2022. [Online]. Available: https://xra.org/research-best-practices/

[413] R. Metz. (2022). Meta is giving avatars a four-foot buffer zone to cut back on VR harassment. CNN Business, Cable News Network, A Warner Media Company. Accessed: Mar. 25, 2022. [Online]. Available: https://edition.cnn.com/2022/02/04/tech/meta-vr-avatar-safety-boundary/index.html

[414] M. Slater, C. Gonzalez-Liencres, P. Haggard, C. Vinkers, R. Gregory-Clarke, S. Jelley, Z. Watson, G. Breen, R. Schwarz, W. Steptoe, D. Szostak, S. Halan, D. Fox, and J. Silver, "The ethics of realism in virtual and augmented reality," *Frontiers Virtual Reality*, vol. 1, pp. 1–13, Mar. 2020.

[415] S. Dijkstra-Soudarissanane, T. Klunder, A. Brandt, and O. Niamut, "Towards XR communication for visiting elderly at nursing homes," in *Proc. ACM Int. Conf. Interact. Media Exper.* New York, NY, USA: Association for Computing Machinery, 2021, pp. 319–321, doi: 10.1145/3452918.3467815.

[416] C. Ziker, B. Truman, and H. Dodds, *Cross Reality (XR): Challenges and Opportunities Across the Spectrum*. Cham, Switzerland: Springer, 2021, pp. 55–77, doi: 10.1007/978-3-030-58948-6_4.

[417] N. de la Peña, P. Weil, J. Llobera, B. Spanlang, D. Friedman, M. V. Sanchez-Vives, and M. Slater, "Immersive journalism: Immersive virtual reality for the first-person experience of news," *Presence, Teleoperators Virtual Environ.*, vol. 19, no. 4, pp. 291–301, Aug. 2010, doi: 10.1162/PRES_a_00005.

[418] S. Hussain, *Multisensory Digital Experiences: Integrating New Interactive Technologies With Human Senses*. Hershey, PA, USA: IGI Global, 2021, pp. 371–386.

[419] P. Sykownik and M. Masuch, "The experience of social touch in multiuser virtual reality," in *Proc. 26th ACM Symp. Virtual Reality Softw. Technol.* New York, NY, USA: Association for Computing Machinery, 2020, Art. no. 30, doi: 10.1145/3385956.3418944.

[420] C. Peter, A. Kreiner, M. Schröter, H. Kim, G. Bieber, F. Öhberg, K. Hoshi, E. L. Waterworth, J. Waterworth, and S. Ballesteros, "AGNES: Connecting people in a multimodal way," *J. Multimodal User Interfaces*, vol. 7, no. 3, pp. 229–245, Nov. 2013. [Online]. Available: http:%5CLibrary%5CPeterKreiner2013

JANTO SKOWRONEK received the doctoral degree (Dr.-Ing.) from TU Berlin, Germany, in 2016, and the Diplom-Ingenieur degree in electrical engineering and information sciences from Ruhr-University Bochum, Germany. Before his Ph.D. research, he worked at the Digital Signal Processing Group, Philips Research, Eindhoven, The Netherlands, from 2003 to 2010. His main topics at that time were auditory and multi-sensory perception and its applications in digital signal processing algorithms as well as audio and music content analysis using machine learning techniques. He is currently the Managing Director of the research thrust "Smart Technologies, Processes, and Methods" at the Hochschule für Technik Stuttgart–University of Applied Sciences, Stuttgart, Germany. In this role, he is responsible for both strategic development and background operations of the research thrust in which about 30 professors and about 30 research staff members are active. Moreover, he is also coordinating a development team that is creating a transfer platform for the university and he is engaged in various activities for improving work processes in the university with digital tools. Especially for the latter role, he draws from his scientific expertise on quality of experience of telemeetings, which he built up during his tenure as a Research Assistant at TU Berlin and a Research Assistant and a Postdoctoral Researcher at TU Ilmenau, Germany, from 2015 to 2018. Since 2012, he has been a Co-Rapporteur of Question 10 "Conferencing and telemeeting assessment" of the ITU-T Study Group 12 "Performance, QoS, and QoE."

ALEXANDER RAAKE (Member, IEEE) received the doctoral degree (Dr.-Ing.) from the Electrical Engineering and Information Technology Faculty, Ruhr-Universität Bochum, in January 2005, with a book on the speech quality of VoIP—*Speech Quality of VoIP* (Wiley, 2006). Before he received the Ph.D. degree, he studied electrical engineering and physics at RWTH, Aachen, and ENST/Télécom, Paris, with a subsequent research stay at EPFL, Lausanne, Switzerland. From 2004 to 2005, he was a Postdoctoral Researcher at LIMSI-CNRS, Orsay, France. From 2005 to 2009, he was a Senior Scientist at the Quality and Usability Laboratory of T-Labs, TU Berlin. From 2009 to 2015, he held Assistant Professor and then Associate Professor positions at TU Berlin, heading the Assessment of IP-Based Applications Group at TU Berlin's An-Institut T-Labs, a joint venture between Deutsche Telekom AG and TU Berlin. He was appointed as the Head of the Audiovisual Technology Group and a Full Professor at TU Ilmenau, in 2015. In 2021, he was appointed as the Director of the Institute for Media Technology. Since 1999, he has been involved in the standardization activities of the International Telecommunication Union (ITU-T) on performance, quality of service (QoS), and quality of experience (QoE), where he acts as a Co-Rapporteur for question Q.14/12 on monitoring models for audiovisual services. His research interests include speech, audio, and video communication, quality of experience, audiovisual and multimedia services and networks, and human perception and cognition.

GUNILLA H. BERNDTSSON received the M.Sc. degree in engineering physics from the Chalmers University of Technology, Gothenburg, Sweden, in 1988, and the Ph.D. degree from the Department of Speech Communication and Music Acoustics, Royal Institute of Technology, Stockholm, in 1995. Since 1997, she has been a Senior Researcher at Ericsson Research, Stockholm. She currently works at the Research Department Digital Representations and Interaction, leading the interaction research. Her main focus is on Quality of Experience of eXtended Reality (XR) meetings. She is active in the International Telecommunication Union lead Study Group on performance, Quality of Service (QoS), and Quality of Experience (QoE). Since 2011, she has been a Co-Rapporteur for ITU-T Study Group 12 Question 10 "Conferencing and telemeeting assessment." Since 2013, she has also been the Vice Chair of ITU-T SG12 Working Party 1 "Terminals and multimedia subjective assessment."

OLLI S. RUMMUKAINEN (Member, IEEE) was born in Joensuu, Finland, in 1986. He received the B.Sc., M.Sc., and D.Sc. degrees in electrical engineering from Aalto University, Espoo, Finland, in 2011, 2012, and 2016, respectively.

Since 2016, he has been working at the International Audio Laboratories Erlangen (a joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg and Fraunhofer IIS), Germany, where he is currently a Senior Scientist. His research interests include spatial audio, psychoacoustics, and quality of experience in six-degrees-of-freedom virtual reality.

Dr. Rummukainen was a recipient of the IEEE International Workshop on Quality of Multimedia Experience Best Student Paper Award, in 2012 and 2013, and the Audio Engineering Society International Conference on Audio for Virtual and Augmented Reality Best Paper Award, in 2018. He contributes actively to the ITU-T Study Group 12, where he is a Co-Editor of a work item on QoE assessment of eXtended Reality (XR) meetings.

PAOLINO USAI received the Ph.D. degree in physics, in 1981. He was an Engineer in telecommunications, in 1969. Formerly, he was a Master Researcher at CSELT, the Research and Development Center of Telecom Italia Group, Turin, Italy. From September 1969 to mid-1997, he was responsible for the Subjective Testing Laboratory, Human Factors Group, Customer Terminals Division. At CSELT, he has developed expertise in electro-acoustics, human factors, perception, speech coding, speech synthesis, and speech recognition. He has been an Active Member of ITU-T, since 1981. Since the beginning, he has been an active participant in ETSI speech coding activities (since the set-up of ETSI, in mid-1980s). He joined ETSI at Sophia Antipolis, France, in July 1997, as a Project Manager for radio and speech aspects, until January 2020. At ETSI, he worked as a member of the Project Team (PT 12) during the standardization of the GSM system, then a member of the Mobile Competence Center (MCC), acting in parallel as a Secretary of 3GPP Technical Specification Group GERAN (GSM/EDGE Radio Access Network), 3GPP TSG-GERAN WG1 (Radio Aspects), and 3GPP TSG System Aspects WG4 (Codec), and the Chairperson of the 3GPP SA4 Speech Quality Sub-Working Group. In ITU-T SG12, during a career lasting more than 40 years, he has been since mid-1990s as the Chairperson of the Speech Quality Experts Group within Study Group 12, covering subjective and objective assessment of analogue and digital networks and terminals and then nominated as a Rapporteur of several questions, at present a Co-Rapporteur of Q. 7/12 "Methods, tools and test plans for the subjective assessment of speech and audio-visual quality." He is the author of more than 100 relevant papers/publications (GLOBECOM, ICC, ICASSP, IEEE COMMAG, and Acoustics and Human Factors conferences and workshops) about the performance evaluation of all ITU-T and ETSI speech coding standards.

SIMON N. B. GUNKEL (Senior Member, IEEE) received the master's degree in computer science from Technical University Berlin, Germany, and KAIST, Daejeon, Republic of Korea, in 2012. He is currently pursuing the Ph.D. degree with the Centrum Voor Wiskunde en Informatica (CWI), Amsterdam, The Netherlands. He is a Research and Development Scientist at TNO, the Netherlands Organization for Applied Scientific Research, The Hague. His current focus is on the topic of web-based photo-realistic social XR, which is also the topic of his ongoing Ph.D. research in collaboration with CWI. His activities cover the full innovation cycle from ideation, implementation, evaluation, and standardization (i.e., 3GPP SA4). Part of his work includes leading the technical work package in the Horizon 2020 Project VRTogether (http://vrtogether.eu/). Furthermore, he is holding a board position as a Technology Manager at the Young European Associated Researchers Network—YEAR (http://www.year-network.com/). His ongoing efforts in applied research cover a wide range of topics in multimedia, such as virtual reality communication, video conferencing, and quality of experience.

MATHIAS JOHANSON was born in Gothenburg, Sweden, in 1971. He received the M.Sc. degree in computer science and the Ph.D. degree from the Chalmers University of Technology, in 1994 and 2003, respectively. He is currently the Co-Founder and the Research and Development Manager of Alkit Communications AB, where he for many years has led the development of video communication products and services. In addition to his expertise in video communication, his professional interests also include automotive telematics and intelligent transportation systems. He has participated in many European and national research and development projects in collaboration between industry and academia. He has authored more than 40 peer-reviewed papers and one book chapter.

EMANUËL A. P. HABETS (Senior Member, IEEE) was born in Maastricht, The Netherlands, in 1976. He received the B.Sc. degree in electrical engineering from the Hogeschool, Limburg, The Netherlands, in 1999, and the M.Sc. and Ph.D. degrees in electrical engineering from Technische Universiteit Eindhoven, The Netherlands, in 2002 and 2007, respectively.

From 2007 to 2009, he was a Postdoctoral Fellow at the Technion—Israel Institute of Technology and Bar-Ilan University, Israel. From 2009 to 2010, he was a Research Fellow with the Communication and Signal Processing Group, Imperial College London, U.K. He is currently an Associate Professor at the International Audio Laboratories Erlangen (a joint institution of the Friedrich-Alexander-Universitat Erlangen-Nurnberg and Fraunhofer IIS, Germany) and the Head of the Spatial Audio Research Group at Fraunhofer IIS. His research interests include audio and acoustic signal processing, spatial audio signal processing, spatial sound recording and reproduction, speech enhancement (dereverberation, noise reduction, echo reduction), and sound localization and tracking.

Dr. Habets is also a member of the EURASIP Technical Activities Board and the Chair of the EURASIP Technical Area Committee on Acoustic, Speech and Music Signal Processing. He was a recipient, with S. Gannot and I. Cohen, of the 2014 IEEE Signal Processing Letters Best Paper Award.

LUDOVIC MALFAIT received the master's degree in telecommunication from Telecom Lille, France, in 2001, and the Ph.D. degree in computer science from Essex University, U.K., in 2010.

From 2001 to 2012, his research at Psytechnics focused on the objective and subjective assessment of voice and video communication systems. Since 2012, his work at Dolby has been on the evaluation of spatial audio telemeetings and more recently on advanced media delivery. He is currently a Research and Development Engineer at Dolby Laboratories. Since 2001, he has been involved in standardization activities at ITU-T Study Group 12, the group responsible for standards on performance, quality of service (QoS), and quality of experience (QoE). He was appointed as the Vice-Chair for Working Party 2 "Objective models and tools for multimedia quality," and he is a Co-Rapporteur for Question 7 "Methodologies, tools and test plans for the subjective assessment of speech, audio and audiovisual quality interactions."

DAVID LINDERO was born in Skellefteå, Sweden, in 1981. He received the M.Sc. degree in signal processing from the Luleå University of Technology (LTU), in 2007.

Since 2007, he has been working at Ericsson Research, Luleå. Based on the focus on psychoacoustics in the M.Sc. studies, the initial topics were within the field of audio- and video communication service quality. This work has since then also been branching out into areas, such as cloud gaming, remotely rendered VR, and generic AI/machine learning; always keeping the Quality of Experience context. In 2015–2016, he did a one-year project as a Research Associate at LTU, investigating the effects of digitally encoded and compressed speech on cognitive load. As part of his research work, he has also been working actively with the ITU-T Study Group 12 within many of the questions. Both participating not only in modeling projects to create objective assessment methods but also in developing the common statistical evaluation methods and other tools that contribute to many activities with the ITU-T.

ALEXANDER TOET (Life Senior Member, IEEE) was born in Apeldoorn, The Netherlands, in 1955. He received the Ph.D. degree in physics from the University of Utrecht, The Netherlands, in 1987. He worked on visual–spatial localization (hyperacuity) and digital image processing at the University of Utrecht. He is currently a Senior Research Scientist at TNO, Soesterberg, The Netherlands, where he investigates multisensory (visual, auditory, tactile, and olfactory) perception and the crossmodal interactions between these different senses, with the aim to optimize the quality of user experience of virtual, augmented, and mixed reality environments. He is affiliated with the Faculty of Social and Behavioral Sciences, University of Utrecht, where he researches the visual perception of fused multimodal imagery. He also investigates to what extent multisensory telerobotic system interfaces can elicit the senses of telepresence and embodiment, with the aim to enhance task performance in remote operations. His current focus is on the development of (subjective and objective) tools to assess the experienced quality of mediated social communication systems. He is a fellow of the International Society for Optical Engineering (SPIE) and a member of the SAE-10 Technical Committee on Laser Safety Hazards.