University of Zagreb

FACULTY OF ELECTRICAL ENGINEERING AND COMPUTING

Dunja Vučić

QUALITY OF EXPERIENCE DRIVEN VIDEO ENCODING ADAPTATION FOR MULTIPARTY AUDIOVISUAL TELEMEETINGS ON MOBILE DEVICES

DOCTORAL THESIS

Zagreb, 2021.

University of Zagreb

FACULTY OF ELECTRICAL ENGINEERING AND COMPUTING

Dunja Vučić

QUALITY OF EXPERIENCE DRIVEN VIDEO ENCODING ADAPTATION FOR MULTIPARTY AUDIOVISUAL TELEMEETINGS ON MOBILE DEVICES

DOCTORAL THESIS

Supervisor: Professor Lea Skorin-Kapov, PhD

Zagreb, 2021

Sveučilište u Zagrebu

FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

Dunja Vučić

PRILAGODBA KODIRANJA VIDEA VOĐENA POBOLJŠANJEM ISKUSTVENE KVALITETE VIŠEKORISNIČKIH AUDIOVIZUALNIH DALJINSKIH SASTANAKA NA POKRETNIM UREĐAJIMA

DOKTORSKI RAD

Mentorica: prof. dr. sc. Lea Skorin-Kapov

Zagreb, 2021.

The doctoral thesis was completed at the University of Zagreb Faculty of Electrical Engineering and Computing, Department of Telecommunications.

Supervisor: Professor Lea Skorin-Kapov, PhD

This thesis has: 208 pages

Thesis number: ___________

About the Supervisor

Lea Skorin-Kapov is Professor and head of the Multimedia Quality of Experience Research Lab (MUEXLab) at the University of Zagreb Faculty of Electrical Engineering and Computing. Her research interests include Quality of Experience modeling of advanced multimedia applications, QoE monitoring of encrypted video traffic, and cross-layer negotiation and management of QoS/QoE in networks. She teaches courses at bachelor, masters, and doctoral levels dealing with multimedia services, heuristic optimization methods, and communication networks.

She received her Dipl.-Ing., M.S., and Ph.D. degrees in Telecommunications from the Faculty of Electrical Engineering and Computing (FER) at the University of Zagreb, Croatia, in 2001, 2004, and 2007, respectively. From 2001-2009 she was employed in the Research and Development Center of Ericsson Nikola Tesla d.d. (ETK), Zagreb, Croatia, doing research on QoS signaling, negotiation, and adaptation for multimedia services. From 2002-2009 she was also an adjunct teaching and research assistant at the Department of Telecommunications, FER, University of Zagreb.

Since 2010 she has been employed at the Department of Telecommunications, FER, University of Zagreb. She has been involved in a number of industry and research projects, and is currently principal investigator for the project “Modeling and Monitoring QoE for Immersive 5G-Enabled Multimedia Services”, funded by the Croatian Science Foundation. She has published over 100 scientific papers, and serves on the editorial boards of IEEE Transactions on Network and Service Management and Springer’s Multimedia Systems journal, and has served as Guest Editor for the IEEE Journal of Selected Topics in Signal Processing, and ACM Transactions on Multimedia Computing, Communications, and Applications.

Lea Skorin-Kapov is a senior member of IEEE and is currently serving as Chapter chair of the IEEE Communications Society - Croatia Chapter.

O mentorici

Lea Skorin-Kapov je redovita profesorica na Zavodu za telekomunikacije Fakulteta elektrotehnike i računarstva (FER) Sveučilišta u Zagrebu te voditeljica istraživačkog laboratorija Multimedia Quality of Experience Research Lab (MUEXLab). Njezino glavno područje istraživačkog interesa jest modeliranje iskustvene kvalitete višemedijskih usluga, praćenje iskustvene kvalitete te mehanizmi upravljanja i optimizacije kvalitete usluge/ iskustvene kvalitete u mrežama. Podučava na FER-ovom preddiplomskom, diplomskom i doktorskom studiju o višemedijskim uslugama i komunikacijama, heurističkim metodama optimizacije i komunikacijskim mrežama.

Diplomirala je 2001. godine, magistrirala 2004. godine te doktorirala 2007. godine na Sveučilištu u Zagrebu FER, smjer telekomunikacije i informatika. Od 2001.-2009. godine radila je u Istraživačkom odjelu tvrtke Ericsson Nikole Tesle d.d. (ETK), Zagreb u jedinici za Istraživanje i razvoj, gdje se bavila istraživanjem područja signalizacije, pregovaranja i prilagodbe kvalitete usluge za napredne višemedijske usluge. Od 2002. do 2009. godine radila je kao mlađi asistent na Zavodu za telekomunikacije na FER-u, Sveučilišta u Zagrebu.

Od 2010. godine zaposlena je na Zavodu za telekomunikacije, FER, Sveučilište u Zagrebu. Uključena je u razne istraživačke i industrijske projekte. Trenutno vodi istraživački projekt "Modeliranje i praćenje iskustvene kvalitete imerzivnih višemedijskih usluga u 5G mrežama" kojeg financira Hrvatska zaklada za znanost. Autorica je i koautorica više od 100 znanstvenih radova. Članica je uredničkog odbora časopisa IEEE Transactions on Network and Service Management i Springerovog časopisa Multimedia Systems, te je sudjelovala kao gostujući urednik za časopise IEEE Journal of Selected Topics in Signal Processing i ACM Transactions on Multimedia Computing, Communications, and Applications.

Trenutno obnaša funkciju predsjednice Odjela za komunikacije, hrvatske sekcije IEEE.

Acknowledgements

First and foremost I am extremely grateful to my supervisor Prof. Lea Skorin-Kapov for her invaluable advice, and continuous support during my PhD study. Her knowledge and plentiful experience have encouraged me in all the time of my academic research. I would also like to thank all the members at the Department of Telecommunications, Faculty of Electrical Engineering and Computing, and colleagues from Ericsson Nikola Tesla.

Special thanks goes to the students of the Faculty of Electrical Engineering and Computing, Tima Redžović, Magdalena Lisica, Tina Knežević, and Ivona Mihotić, for their contribution in helping, organizing and conducting subjective user studies.

Finally, I would like to express my gratitude to my beloved family and best friends. Without their support and encouragement, it would be impossible for me to complete my study.

Abstract

Video conferencing is becoming increasingly popular in both leisure and business contexts, offering opportunities to communicate with family, friends, and colleagues, increase productivity, reduce costs, and share information in real time. High resolution displays, front and rear cameras, high speed mobile networks and modern technologies, such as WebRTC (Web Real-Time Communication), are contributing to making video conferencing free and available "anywhere at any time". However, given the strict delay and high bandwidth requirements associated with video conferencing services, along with variable mobile network resource availability and limited mobile end user device capabilities, dynamic service adaptation strategies are needed to achieve acceptable end-user perceived quality. The main objective of this research is to identify and quantify the impact of various encoding, system, and network influence factors on Quality of Experience (QoE) during multiparty audiovisual telemeetings on mobile devices, with the aim to work towards QoE-driven service adaptation strategies.

Understanding and modeling the QoE of audiovisual telemeeting services is a complex multi-layered problem, whereby numerous factors impacting QoE and related to the system, context or user make it difficult to obtain reliable models and interpret results. Therefore QoE management approaches generally do not focus on a single factor, but rather need to consider a combination of factors and their joint impact on QoE. To specify key system-, contextual-, and human influence factors that impact QoE and corresponding QoE dimensions in the context of multiparty audiovisual telemeetings on mobile devices, we conducted an online survey in order to gather user feedback (reported by 272 participants). Identified factors can be used as a predictors when modeling QoE and enhance model accuracy. However, besides higher complexity, models with a large number of predictors can suffer from the problem of overfitting and can be hard to interpret, especially when predictors are correlated with each other. Hence, a good balance between accuracy and complexity has to be found.

In this thesis, we present the results of six conducted empirical subjective user studies in a leisure context that investigate the impact of system, network, and video encoding parameters (namely video bitrate, resolution, and frame rate) on perceived quality for multiparty audiovisual telemeetings on mobile devices. Different test conditions based on the video encoding parameters were rated (in terms of perceived overall, audio, and video quality) during experiments, and provided the input for proposing QoE and perceived video quality estimation models. The proposed QoE model quantifies the relationship between QoE and perceived video and audio quality, while the perceived video quality model quantifies the relationship between objective (in terms of video encoding parameters, and in terms of blurriness and blockiness) and subjective quality. Based on the derived models, we proposed QoE-driven video encoding adaptation strategies for multiparty audiovisual telemeetings on mobile devices, designed to ensure satisfactory QoE under variable system and network resource availability constraints.

Key words: multiparty telemeetings, Quality of Experience, user studies, video encoding parameters, adaptation strategies, mobile devices

Prošireni sažetak

Glavni cilj ovog istraživanja je specifikacija strategija prilagodbe video kodiranja kako bi se optimizirala iskustvena kvaliteta krajnjih korisnika usluge višekorisničkog audiovizualnog daljinskog sastanka, ostvarenog putem pokretnih uređaja uz ograničenja raspoloživih resursa. Kroz šest korisničkih studija provedenih između 2015. i 2018. godine. istražen je utjecaj parametara video kodiranja (rezolucija, brzina kodiranja i brzina okvira) te mreže (gubitak paketa i kašnjenje) na iskustvenu kvalitetu.

U posljednja dva desetljeća video usluge korištene putem Interneta doživjele su značajan porast, što je bilo omogućeno tehnološkim napretkom mreža, sve većim brzinama prijenosa, poboljšanim metodama kodiranja videa i dostupnim krajnjim pokretnim uređajima s kvalitetnim zaslonom, kamerom, zvučnikom i mikrofonom. Pokretni uređaji, usluge i aplikacije postali su dio naše svakodnevice, mijenjajući odnose, društvene norme, metode komunikacije i načine interakcije. Životne okolnosti i ubrzani način života stvorili su potrebu za komunikacijom putem pokretnih uređaja čak i prije izbijanja COVID-19 pandemije. Video pozivi su postali popularne i neizostavne aplikacije kako u poslovnom tako i u privatnom kontekstu.

Prije nekoliko godina videokonferencijski sustavi bili su većinom namijenjeni poslovnoj upotrebi unutar velikih organizacija koje su si mogle priuštiti konferencijske sobe (kao krajnje točke sustava), infrastrukturu i osoblje potrebno za implementaciju i održavanje tako složenih sustava. S vremenom su krajnje točke sustava postala široko rasprostranjena stolna računala, koja su i dalje uglavnom bila dostupna velikim organizacijama. Kako se povećavao broj krajnjih točaka, tako je potrebna infrastruktura postajala sve složenija, što je u konačnici rezultiralo i višim operativnim troškovima. Posljednjih godina se pojavila mogućnost smanjenja operativnih troškova uporabom rješenja zasnovanih u oblaku, gdje pružatelj usluge osigurava i potrebnu infrastrukturu. Tehnologije poput WebRTC-a (Web Real-Time Communications), omogućile su besplatne pozive velikom broju korisnika pokretnih uređaja bilo kad i bilo gdje. WebRTC je projekt otvorenog koda, koji omogućava razvoj aplikacija za video pozive i razmjenu podataka između preglednika bez potrebe za instalacijom programskih dodataka. Korisnik koji pokreće video poziv stvara online virtualnu sobu putem web aplikacije za pokretanje WebRTC sesije. Ostali korisnici pozivaju se da pristupe virtualnoj sobi putem svog mrežnog preglednika i generiranog lokatora sadržaja. Prije nego što pristupi virtualnoj sobi svaki korisnik mora odobriti pristup kameri i mikrofonu.

Danas su krajnje točke pokretni uređaji, s procesorskim mogućnostima dostatnim za simultano kodiranje i dekodiranje videa pri visokoj prostornoj i vremenskoj rezoluciji tijekom komunikacije u stvarnom vremenu. Iako je svaka nova generacija pametnih telefona naprednija od prethodne generacije, veličina zaslona kod većine pametnih telefona ostala je ispod 6". Izgled aplikacije i raspored elemenata na zaslonu, kao i njihova veličina trebao bi se prilagoditi

veličini i orijentaciji pojedinog pokretnog uređaja. Međutim, u slučaju video poziva s tri ili više korisnika, posebno kada se radi o slučaju da su svi korisnici poziva prikazani na zaslonu istovremeno, veličina prozora pojedinog korisnika je mala, što omogućava smanjenje video rezolucije bez značajnog utjecaja na iskustvenu kvalitetu.

Korisnici očekuju da će usluga video poziva (uspostavljenog u različitim kontekstima) biti pouzdana i dostupna kroz heterogene pristupne mreže i uređaje. Korištene tehnologije takvih naprednih aplikacija moraju biti sigurne i jednostavne za upravljanje. Bez obzira na složenost sustava, sama usluga mora biti jednostavna i korisnici bi je trebali moći koristiti bez intenzivnog treninga. Aplikacije moraju podržavati dodatne funkcionalnosti i alate koji su posebno važni za poslovni kontekst kako bi se omogućila bolja suradnja. Tako je mogućnost dijeljenja sadržaja u realnom vremenu jedna od ključnih značajki interaktivnih sastanaka. Sudionici daljinskih sastanaka bi trebali moći i snimiti sastanak, spremiti ga i nakon toga ga jednostavno podijeliti.

Video pozivi koji se koriste u poslovnom kontekstu općenito imaju definirani cilj, određen nizom zadataka koje je potrebno izvršiti. S druge strane, osnovni cilj korištenja video poziva u privatnom kontekstu odnosno u slobodno vrijeme jest doživjeti osjećaj prisutnosti ili društvene povezanosti. Zbog različitih ciljeva sastanaka ostvarenih u poslovnom i privatnom kontekstu, očekivana percipirana kvaliteta usluge može biti različita, pri čemu će sudionici vjerojatno biti manje kritični kada je u pitanju privatni kontekst. Stoga dizajn i specifičnosti implementacije audiovizualnih daljinskih sastanaka upravo i ovise o kontekstu, broju sudionika i njihovim potrebama. Dizajneri moraju uzeti u obzir i hardverske mogućnosti krajnjih uređaja poput mikrofona i zvučnika odnosno veličine zaslona i kvaliteta kamere. Međutim, iako hardver može snimiti video visoke kvalitete, to ne znači nužno da će procesorske mogućnosti krajnjih uređaja biti dovoljne za istovremenu obradu više medijskih tokova.

Daljinski višekorisnički audiovizualni sastanci se razlikuju u nekoliko važnih aspekata. Razliku čini broj sudionika, lokacija kao i raspored sudionika po lokacijama. Tako je moguće da se na jednoj lokaciji nalazi više od jedne osobe ili da imamo isti broj lokacija i sudionika. Daljinski sastanci se mogu dodatno razlikovati u pogledu interaktivnosti. Neinteraktivna kvaliteta se može evaluirati samo slušanjem ili gledanjem unaprijed snimljenih sadržaja, dok interaktivnu kvalitetu obično ocjenjuju sudionici koji su i sami uključeni u razgovor. Postav eksperimenta i samo ocjenjivanje se može odvijati u laboratorijskom okruženju ili prirodnom okruženju koje predstavlja situaciju iz stvarnog života. Uvjeti u kojima se izvodi eksperiment mogu biti kontrolirani ili nekontrolirani, pri čemu bi se nekontrolirano okruženje trebalo dobro opisati. Osim toga, važno je razmotriti i način na koji je postavljen cijeli sustav u kontekstu heterogenih uređaja i pristupnih mreža, odnosno radi li se o simetričnom ili nesimetričnom postavu, koji je češći u scenarijima iz stvarnog života. Asimetrični postav može dovesti do toga da sudionici različito percipiraju nastala oštećenja i degradaciju kvalitete, ali i do toga da imaju različita očekivanja. Kao primjer, možemo zamisliti scenarij iz stvarnog života koji uključuje dva sudionika s vrhunskim računalima i velikim zaslonima povezanim na brze fiksne mreže, koji komuniciraju s trećim sudionikom koji koristi pametni telefon s 5.1" zaslonom, putuje vlakom i ima lošu konekciju s pokretnom mrežom.

Ta raznolikost u okolini, postavu, kontekstu, otežava definiranje generalizirane metode za procjenu iskustvene kvalitete višekorisničkih daljinskih sastanaka za sve vrste opreme korištene u različitim okolnostima. Stoga je u većini naših istraživanja fokus bio na simetričnom postavu u kontroliranim uvjetima (laboratorijskim i kućnim). Kako svaka odluka o postavu eksperimenta utječe na percepciju kvalitete, zadnji aspekt koji je bitno uključiti jest tip zadatka koji će korisnici rješavati tijekom razgovora. Postoji nekoliko standardom definiranih zadataka, međutim rješavanje zadataka tijekom konverzacije može dovesti do toga da sudionici ne gledaju cijelo vrijeme zaslon ukoliko je potrebno koristiti papir i olovku, kao i do toga da njihov angažman pri rješavanju zadataka može utjecati na njihovo ocjenjivanje. Za ocjenjivanje percipirane kvalitete je dozvoljena i poželjna uporaba obične konverzacije bez ikakvih zadataka, međutim ponekad je teško potaknuti i zadržati neprekinutu i glatku konverzaciju među sudionicima koji se ne poznaju od ranije, naročito ako su sudionici sramežljivi ili povučeni. Stoga su u našim eksperimentima grupu uvijek činili sudionici koji se međusobno poznaju. Svi eksperimenti su bazirani na tri sudionika, pri čemu je svaki sudionik bio smješten u jednoj prostoriji. Isto tako, sudionicima je bilo dozvoljeno da sami proizvoljno odaberu na kojoj udaljenosti će biti pametni telefon, odnosno hoće li ga držati u ruci ili na stalku. Sudionicima je na početku objašnjeno i pokazano što se očekuje, provedeno je inicijalno testiranje s ciljem upoznavanja sudionika sa zadatkom i upitnikom koji su trebali popuniti. Preliminarni rezultati nisu uzeti u obzir. Ukupno vrijeme testiranja uvijek treba biti razumno. Kako bi spriječili umor sudionika eksperimenti su bili ograničeni na najviše jedan sat, s pauzom od pet minuta između svakog testnog scenarija.

Video poziv je usluga osjetljiva na kašnjenja i zahtjeva veliku propusnost. Dinamička dostupnost resursa pokretne mreže kao i ograničene mogućnosti pokretnih uređaja krajnjih korisnika, nameću potrebu za prilagodbom usluga kako bi se postigla prihvatljiva razina percipirane kvalitete od strane korisnika. Cilj ovog istraživanja je identificirati i kvantificirati utjecaj različitih faktora (vezanih uz proces kodiranja, sustav i mrežu) na iskustvenu kvalitetu video poziva uspostavljenog putem pokretnih uređaja, kako bi se razvile strategije temeljene na prilagodbi kodiranja videa vođenoj poboljšanjem iskustvene kvalitete. Nastoje se izbjeći nepotrebno visoke brzine kodiranja, brzine okvira kao i video rezolucije koje ne mogu više pridonijeti boljoj iskustvenoj kvaliteti, ali mogu dovesti do zagušenja i zamrzavanja usluge. Korisnici žele imati pristup i pouzdano koristiti zahtjevne usluge bez obzira na kontekst ili faktore koji utječu na percipiranu kvalitetu, poput lokacije, vremena, mrežnih uvjeta ili karakteristika uređaja kojeg koriste. Pokretni uređaji krajnjih korisnika, poput pametnih telefona koji su bili korišteni u studijama mogu predstavljati uska grlo u lancu usluge. Međutim, nova generacija uređaja donosi napredniji hardver u kontekstu radne memorije, snage procesora, kamere i trajanja baterije. Veličina i kvaliteta zaslona kod pametnih telefona se također povećavala kroz vrijeme, tako je 2014. godine rezolucija 1080x1920 px bila implementirana samo u vrhunske pametne telefone. Pet godina kasnije postaje standardna i najčešće korištena rezolucija kod pametnih telefona.

Razumijevanje i modeliranje iskustvene kvalitete za uslugu video poziva ostvarenog putem pokretnih uređaja je višeslojni problem, pri čemu brojni faktori koji utječu na iskustvenu kvalitetu, a povezani su sa sustavom, kontekstom ili korisnikom otežavaju dobivanje pouzdanih modela i tumačenje rezultata. Pri upravljanju iskustvenom kvalitetom nije dobro usredotočiti se samo na jedan faktor, već je potrebno uzeti u obzir kombinaciju faktora i njihov zajednički utjecaj na ukupnu percipiranu kvalitetu. Kako bi se odredili ključni faktori pojedine kategorije: sustav, kontekst i čovjek te odgovarajuće dimenzije iskustvene kvalitete u kontekstu video poziva s više korisnika ostvarenog putem pokretnih uređaja, provedena je anketa putem Interneta. U anketi je sudjelovalo 272 ljudi s ciljem prikupljanja informacija o stavovima i mišljenjima vezanim uz video poziv. Faktori identificirani anketom mogu se koristiti kao nezavisne varijable (prediktori) pri modeliranju iskustvene kvalitete. Upitnik uključuje pitanja kojima se ocjenjuje utjecaj i važnost razmatranih faktora vezanih uz aplikaciju, resurse i kontekst. Odabrani faktori definiraju značajke kvalitete koje šira publika razumije i može ocijeniti. Pitanja su bila podijeljena u četiri skupine: opće podatke (odnosi se na demografske podatke ispitanika), kvaliteta medija (kvaliteta slike i zvuka u kontekstu percipiranih oštećenja), kvaliteta usluge i upotrebljivost (npr. jednostavnost upotrebe i efikasnost) te funkcionalnosti (dodatne funkcionalnosti omogućene povrh samog video poziva, poput dijeljenja datoteka i tekstualnog dopisivanja). Dva su glavna aspekta vezana za pružanje usluge koja se razmatraju pri prikupljanju povratnih informacija: uspostava poziva i način rada same usluge nakon uspostave poziva, odnosno za vrijeme trajanja poziva. Oba aspekta uključuju više dimenzija koje imaju utjecaj na percipiranu iskustvenu kvalitetu, npr. napor koji je potrebno uložiti pri korištenju usluge, dostupnost usluge, točnost prenesenih informacija ili sigurnost. Sljedećih dvanaest faktora je identificirano kao oni koji imaju najviše utjecaja na percipiranu kvalitetu: razumljivost govora, audio-video sinkronizacija, zamrznuti video (duže od 15 sekundi), primjetno kašnjenje zvuka, niska potrošnja baterije, zamućena slika, cijena, sigurnost u kontekstu privatnosti, jednostavno korištenje usluge, primjetno kašnjenje videa, neprekinuta interakcija i složenost instalacije.

Pored ankete, u radu su predstavljeni rezultati korisničkih studija postavljenih u privatnom kontekstu, kojima se istražuje utjecaj parametara sustava, mreže i video kodiranja (poput brzine kodiranja, rezolucije i brzine okvira) na iskustvenu kvalitetu video poziva s tri korisnika, uspostavljenog putem pametnih telefona. Sudionici su ocjenjivali percipiranu audio, video i ukupnu kvalitetu u različitim testnim scenarijima, a dobiveni podatci su korišteni za modeliranje procjene iskustvene kvalitete i percipirane video kvalitete. Predložen model iskustvene kvalitete kvantificira odnos temeljen na percipiranoj kvaliteti zvuka i videa, dok se model za procjenu percipirane video kvalitete temelji na parametrima video kodiranja. Razvijeni modeli služe kao podloga za definiranje strategije prilagodbe višekorisničkih audiovizualnih daljinskih sastanaka na pokretnim uređajima. Strategije su osmišljene tako da se video kvaliteta prilagodi mrežnoj okolini u kontekstu propusnosti, ali i mogućnostima krajnjih uređaja, dok je reaktivni mehanizam za kontrolu zagušenja *(Google Congestion Control)*, temeljen na gubitku paketa i kašnjenju, implementiran u okviru WebRTC projekta. Proaktivan pristup uključuje strategije prilagodbe temeljene na definiranju postavki video kodiranja koje rezultiraju zadovoljavajućom iskustvenom kvalitetom, pri čemu se nastoje izbjeći situacije koje aktiviraju mehanizam kontrole zagušenja.

Rezultati dobiveni u provedenim istraživanjima korišteni su za definiranje strategije prilagodbe kodiranja videa vođene poboljšanjem iskustvene kvalitete u kontekstu ograničenih resursa sustava i mreže. Za dobru iskustvenu kvalitetu potrebno je pronaći dobar omjer rezolucije i brzine kodiranja, ovisan o procesorskim mogućnostima obrade krajnjeg uređaja, kretanju kamere ili sudionika i dostupnoj propusnosti. Izvedeni modeli iskustvene kvalitete i percipirane video kvalitete te strategije video adaptacije razvijeni su isključivo na temelju subjektivnih ocjena korisnika prikupljenih u studijama, pri čemu se koristio VP8 kodek. U radu su definirane tri strategije prilagodbe. Prva strategija se temelji na izvedenim modelima percipirane kvalitete, pri čemu se nastoji maksimizirati iskustvena kvaliteta u odnosu na dostupnu propusnost i brzinu kodiranja. U ovom slučaju propusnost označava raspoloživi resurs za definiranje ciljane brzine kodiranja. Sljedeća strategija prilagodbe se temelji na unaprijed definiranim razinama video kvalitete što će ovisno o raspoloživim mrežnim resursima omogućiti brzo prebacivanje između visoke, srednje i niske razine video kvalitete. Razine video kvalitete su definirane u skladu s rezultatima provedenih subjektivnih studija. Pri čemu visoka razina video kvalitete uključuje parametre kodiranja (rezoluciju, brzinu video kodiranja, brzinu okvira) koji mogu osigurati vrlo dobru ili izvrsnu iskustvenu kvalitetu (odnosi se na postavke parametara koji su rezultirali srednjim ocjenama višim od 4) u kontekstu tipičnom za višekorisničke audiovizualne daljinske sastanke. Srednja razina video kvalitete bi trebala osigurati dobru (prema vrlo dobroj) percipiranu ukupnu kvalitetu, dok najniža razina video kvalitete ujedno predstavlja i donju granicu kvalitete, koja bi i dalje trebala osigurati prihvatljivu iskustvenu kvalitetu. Razine kvalitete su određene u privatnom kontekstu video sastanka s tri sudionika (koji su istovremeno prikazani na zaslonu) i pametnih telefona s minimalno 3 GB radne memorije. Ovaj pristup se temelji na dostupnoj propusnosti, pri čemu se kvaliteta prilagođava ako procijenjena propusnost nije dovoljna za trenutno postavljenu razinu kvalitete. Treća strategija prilagodbe se temelji na opterećenosti procesora. Ova strategija se ne oslanja na određenu razinu kvalitete, već određuje parametre kodiranja koji pružaju najbolju moguću iskustvenu kvalitetu s obzirom na postav. Prilagodba se oslanja na praćenje opterećenja procesora, pri čemu se video kvaliteta smanjuje dok se ne dostigne definirana razina opterećenosti procesora. Prvo je potrebno utvrditi prihvatljivu rezoluciju u skladu s opterećenjem procesora, a potom i dostatnu brzinu kodiranja. Sve tri navedene strategije predstavljaju moguća prilagođenja koja mogu pridonijeti optimizaciji resursa uz prihvatljivu razinu iskustvene kvalitete u zadanom kontekstu.

Tehnološki napredak neprestano mijenja potrošačke trendove. Pokretni telefoni, koji su nekada služili samo za razgovor i slanje poruka, evoluirali su u pametne telefone koje koristimo kako za razonodu tako i za izvršavanje dnevnih zadataka. Napredniji hardver pametnih telefona i peta generacija pokretnih mreža omogućit će veće brzine, manje kašnjenje i pouzdaniju povezanost te na taj način ispuniti preduvjete za dobru iskustvenu kvalitetu višekorisničkih video poziva.

Contents

Chapter 1

Introduction

This chapter presents the background and motivation for this thesis, gives an overview of the problem definition and method of solution, and summarizes the main research contributions.

1.1 Background and motivation

In the past two decades, video transmission over the Internet has experienced significant rise, enabled by technological advancements such as higher network transmission rates, improved video coding capabilities, and the widespread availability of high quality displays, cameras, speakers and microphones on heterogeneous end user devices. Mobile devices, services, and applications have become an inseparable part of our daily lives, affecting relationships, social norms, communication and interaction methods even before global outbreak of the COVID-19 pandemic. Constantly evolving life dynamics and accelerated lifestyle created the need for audiovisual communication, both in business and private contexts.

According to Statista, in 2019 the number of smartphone users worldwide exceeded three billion, with forecasts for the next few years estimating several hundred million users [1]. In the Sandvine "The Global Internet Phenomena Report COVID-19" authors reported that the pandemic changed the way we use the Internet, which dramatically impacted network usage [2]. Video call applications such as Zoom gained popularity, causing significant video traffic increase from mid-March 2020 onward. While the global video conferencing market size in 2018 was USD 3.02 billion, estimated growth by 2026 was set to USD 6.37 billion [3]. Those mesmerizing numbers create the ground for new innovations and business opportunities.

Several years ago, video conferencing systems were for the most part targeted for business use (e.g., Cisco Webex, Adobe Connect). Modern technologies, such as WebRTC (Web Real-Time Communications), made video conferencing free and available to the wider public. With the processing power of mobile devices such as smartphones and tablets becoming sufficient to simultaneously encode and decode video at a high spatial and temporal resolution during real-time communication, mobile video communication service use has grown rapidly [3].

While each new generation of smartphones is more powerful than previous generations in terms of processing power, the majority of screen sizes remains under 6 inches [4]. It is well known that layout should be able to respond to the display size and orientation, but in the case of multiparty video communications where all participants are displayed, the size of each video preview window is inherently limited. Given such small preview window sizes, it is possible to reduce the video resolution without significantly impacting QoE [5].

Given end user needs and expectations, modern video conferencing services are expected to be reliable, and available across heterogeneous access networks, devices, and usage contexts. Underlying technologies, in terms of platforms and protocols, of such advanced applications need to be secure and easy to manage. Regardless of the system complexity, the service itself has to be simple and participants should be able to use it without intense training. Features and functions must be useful and access seamless, especially for a business context. Video conferencing used in a business context generally has a specific objective, with a set of tasks that must be completed. On the other hand, video conferencing used in the private/leisure context generally has the primary objective to experience a sense of presence or social connection. Due to the different objectives of the meeting, the quality expected by the participants may be different, with participants likely being less critical when it comes to the private context [6], [7].

Going beyond conversational services between two participants, due to the impacts of the COVID-19 pandemic, users are increasingly using multiparty video conferencing services in both business and leisure contexts (e.g., social interactions via Skype, Viber, Whatsapp, Whereby, Zoom, Microsoft Teams, Webex, etc.). Such multiparty settings impose a wide range of challenges with respect to identifying and quantifying the impact of various factors influencing end user QoE.

In the case of multiparty video calls established in mobile environments, we can observe influence factors related to the end user device, network, context and content. A wide range of smartphone models available on the market along with different access networks in real life can create numerous different asymmetric scenarios. Video calls impose strict low latency and high volume requirements on the underlying network. However, variable network conditions imply the need for dynamic service adaptation and optimization mechanisms. In particular, video encoding parameters such as resolution, bitrate, and frame rate can be dynamically adapted to reduce traffic in light of limited bandwidth availability. The challenge lies in determining how to adapt such parameters in a QoE-aware manner, taking into account additional factors such as context, the number of call participants, and mobile device capabilities. To reach acceptable QoE, service providers have to be able to manage and control resources efficiently. Hence, optimization strategies have the task to recognize how much resources will be required in specific scenarios, and what can be adapted to prevent congestion so as to maintain an acceptable level of perceived quality [8].

1.2 Problem statement

Evaluation of a video conferencing service requires assessment of perceived quality by all involved participants. Unidirectional and bidirectional, dyadic, standardized subjective quality assessment methods for several elements used in a video conference, such as speech, codecs, characterized by bitrate (fixed or variable), frame rate, resolution, noise cancellation, background noise, synchronization and transmission impairments are well establish in [9], [10], [11], [12], [13], [14], [10], [11], [15], [16]. ITU-T Recommendation P.1301 on “Subjective quality evaluation of audio and audiovisual multiparty telemeetings” defines the terms necessary for subjective quality assessment of multiparty telemeeting services [17]. However, missing are detailed recommendations multiparty conversational and interactive video service quality assessment in mobile environments and/or focusing on mobile devices.

In the context of mobile networks, characterized by variable network resource availability, challenges arise with respect to meeting the QoE requirements of conversational real-time, media rich, and multi-user services. One way to reduce packet losses and delays resulting from network congestion is to increase bandwidth availability, which imposes increased costs for operators and potentially end users. With the move towards 5G, the aim will be to meet the requirements of low latency and high-volume service scenarios. However, in practice, challenges still remain, in particular in areas with low coverage or very crowded cells. In addition to network requirements, multiparty video conferencing services impose strict requirements in terms of end user device processing capabilities, with the need for real-time encoding and decoding of multiple media streams. To optimize service performance, in particular from a QoE point of view, there is a need for dynamic service adaptation and optimization mechanisms in light of varying resource availability. In particular, the Google Congestion Control (GCC) algorithm is specifically designed to target real-time streams such as telephony and video conferencing. Based on packet loss and bandwidth estimations, the algorithm invokes stream adaptation, including bitrate, resolution, and frame rate adaptation [18], [19].

Our goal may be considered as complementary to such congestion control algorithms (e.g., built into browsers), by focusing on a more proactive approach. The aim is to reduce unnecessarily high volumes of mobile traffic, and prevent mobile equipment overuse, by avoiding unnecessarily high frame rates, video resolutions, and encoding bitrates which cannot contribute to higher QoE, yet may lead to congestion and service freeze. Thus, investigations are needed to determine how devices, application-level parameters, and the network interact with each other and reflect on QoE. A QoE-aware approach to specifying service adaptation algorithms can thus lead to both more efficient use of available resources, as well as enhanced end user QoE.

A comprehensive study of QoE for multiparty conferencing and telemeeting systems providing methods and conceptual models for perceptual assessment and prediction, emphasizing communication complexity and involvement, is given in [20]. In the context of mobile multiparty services, there is a wide range of challenges with respect to identifying and quantifying the impact of various factors influencing end user QoE. In [21], the authors summarize the challenges in properly assessing the QoE of such systems, and highlight mobility aspects, device and encoding interoperability, ease of use, and additional collaboration possibilities (e.g., exchanging pictures, files, chatting). Factors as well as impacts can be different for each participant in a symmetric set-up, but the situation gets more complicated if there are several participants with heterogeneous end user devices and access networks. This diversity of a multiparty system leads to a complex situation calling for extension of well established QoE assessment methods specified for two-party calls. Even in case of only one degradation source, the translation into perceptible impairments is multiplied, and different degradation can lead to different perception [22]. Papers addressing the quality of video conferencing services have to a great extent focused on the perceived quality in desktop environments, with scenarios differing in packet loss, delay, and available bandwidth. Experiments conducted in local networks and mobile or desktop environments evaluating video call quality have shown sensitivity to bursty packet losses and long delays [23], [24], [25], [26], [27], [28], [29], [30].

Quantifying the influence of video resolution, video frame rate and video content type on the QoE by means of objective video quality metrics showed that the quality degradation is smaller for lower resolutions than for lower frame rates. Frame rate decreasing would save less bandwidth and the video experience would be disturbed to a greater extent. Consequently, studies found that releasing bandwidth should be accomplished by reducing the resolution [31]. Bandwidth savings are achieved with various video compression techniques, such as commonly used video coding standards H.264 and VP8, both of which achieve efficient compression and low bitrate [32]. Certain studies comparing H.264 to VP8 showed that H.264 had lower bandwidth usage and better video quality [33], [34].

An important feature of any service is the possibility to adapt the layout and content to viewing contexts and devices. User preferences for single and dual layouts for desktop video conferencing were tested to investigate the relationship between QoE and layout [35], [36]. Different layout/stream configurations were displayed and distortion measurements showed positive correlation to the overall experience. Authors also indicated that audio in some cases has stronger impact than video.

The user's personality can have significant influence on the overall quality, especially in determining the conversational structure, in terms of turn-taking behavior, single-, double- or multi-talk situations. Turn-taking in every day communication usually does not present a problem. However, during a network-based video conversation, participants have been reported having problems identifying the source of impairments as technical (e.g., attributed to network impairments such as delay) or interpersonal, behavior related attributes (e.g., conscientiousness, openness, extroversion, and agreeableness) [37], [38].

Managing interactive and multiparty video conferencing services requires an understanding of the key underlying QoE influence factors. A key challenge faced by multiparty mobile video conference providers lies in configuring the video encoding parameters so as to maximize participant QoE while meeting resource (network and mobile device) availability constraints. Currently developed QoE models can for the most part be applied to two interlocutors and in desktop environments. However, there is a lack of studies that focus on modeling and optimizing QoE for such services when using mobile devices, and in particular in the case of multiparty scenarios. Hence, there is a need to investigate thresholds within video encoding parameters that can be used to determine optimal adaptation strategies for mobile multiparty video conferencing services. We note that the focus of this work is on mobile end user smartphone devices, while the concept of mobility (e.g., standing still, walking, driving) is out of scope of this thesis. However, by investigating QoE in the context of various limited bitrate scenarios, we aim to address the challenge of optimizing multiparty video conferencing services across variable and limited access network bandwidth conditions such as those characteristic of certain mobile network scenarios.

Following a thorough analysis of state of the art work (provided in Chapter 3), the following research questions have been identified to be addressed in the scope of this thesis:

- **RQ1**: What are the most influential factors impacting QoE in the context of mobile multiparty audiovisual telemeetings?
- **RQ2:** How can the relationship between QoE and selected video encoding parameters (bitrate, resolution, frame rate) be quantified for multiparty audiovisual telemeetings established via smartphone devices?
- **RQ3:** Can perceived video quality for multiparty audiovisual telemeetings on mobile devices be estimated based on objective video quality metrics?
- **RQ4:** How can video encoding parameters corresponding to multiparty audiovisual telemeeting services established via smartphone devices be configured so as to optimize end user QoE, given limited processing capabilities of end user mobile devices and bandwidth constraints?
- **RQ5**: What is the impact of packet loss on QoE for multiparty audiovisual telemeetings established via mobile devices?

The research questions are mapped to a set of activities comprising the overall research methodology (Section 1.3), and further to novel scientific contributions provided as the output of this thesis (Section 1.4). This mapping is portrayed in Figure 1.1.

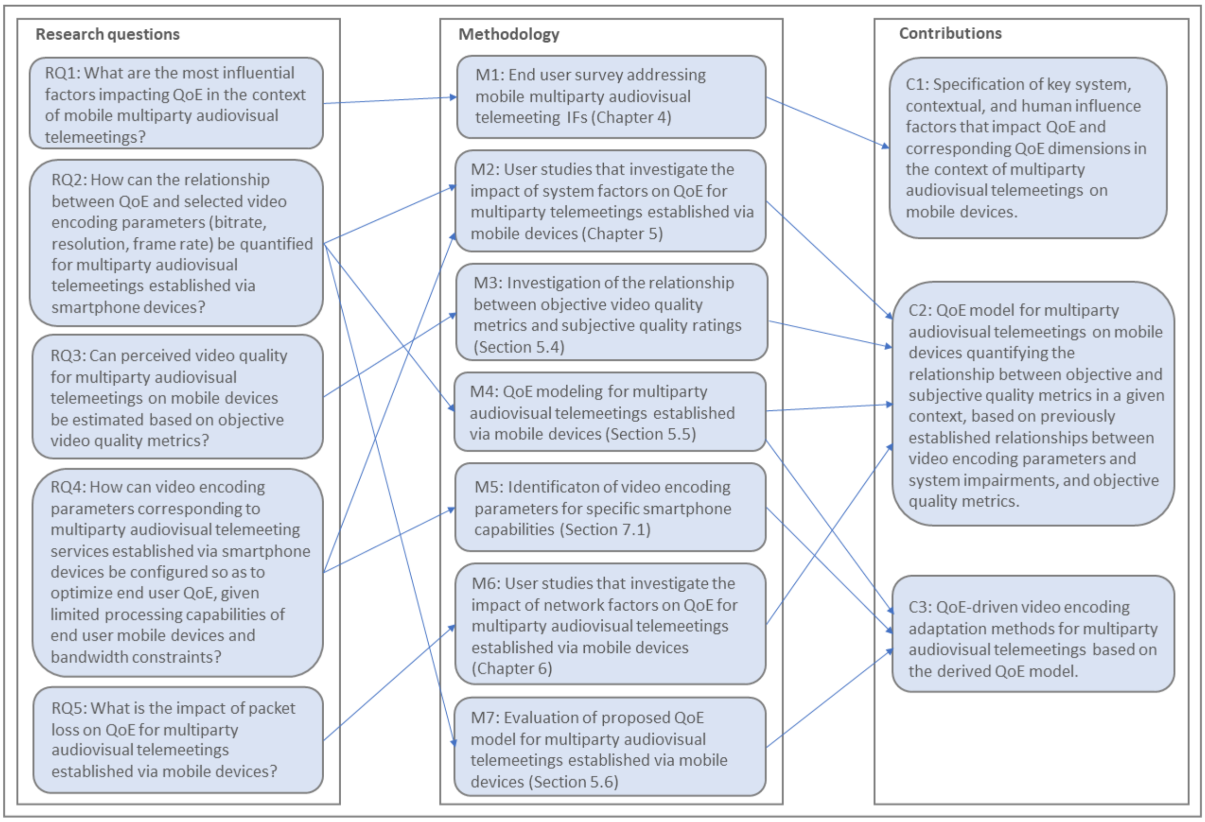

Research questions
RQ1: What are the most influential factors impacting QoE in the context of mobile multiparty audiovisual telemeetings?
RQ2: How can the relationship between QoE and selected video encoding parameters (bitrate, resolution, frame rate) be quantified for multiparty audiovisual telemeetings established via smartphone devices?
RQ3: Can perceived video quality for multiparty audiovisual telemeetings on mobile devices be estimated based on objective video quality metrics?
RQ4: How can video encoding parameters corresponding to multiparty audiovisual telemeeting services established via smartphone devices be configured so as to optimize end user QoE, given limited processing capabilities of end user mobile devices and bandwidth constraints?
RQ5: What is the impact of packet loss on QoE for multiparty audiovisual telemeetings established via mobile devices?
Methodology
M1: End user survey addressing mobile multiparty audiovisual telemeeting IFs (Chapter 4)
M2: User studies that investigate the impact of system factors on QoE for multiparty telemeetings established via mobile devices (Chapter 5)
M3: Investigation of the relationship between objective video quality metrics and subjective quality ratings (Section 5.4)
M4: QoE modeling for multiparty audiovisual telemeetings established via mobile devices (Section 5.5)
M5: Identificaton of video encoding parameters for specific smartphone capabilities (Section 7.1)
M6: User studies that investigate the impact of network factors on QoE for multiparty audiovisual telemeetings established via mobile devices (Chapter 6)
M7: Evaluation of proposed QoE model for multiparty audiovisual telemeetings established via mobile devices (Section 5.6)
Contributions
C1: Specification of key system, contextual, and human influence factors that impact QoE and corresponding QoE dimensions in the context of multiparty audiovisual telemeetings on mobile devices.
C2: QoE model for multiparty audiovisual telemeetings on mobile devices quantifying the relationship between objective and subjective quality metrics in a given context, based on previously established relationships between video encoding parameters and system impairments, and objective quality metrics.
C3: QoE-driven video encoding adaptation methods for multiparty audiovisual telemeetings based on the derived QoE model.

Figure 1.1: Mapping of addressed research questions, methodology, and contributions of the thesis.

1.3 Method of solution and scope

A wide range of parameters influence the overall QoE of multiparty mobile video conferencing services, including various system, user, and context factors. Given the possibilities of different combinations of influence factors (IFs), there is a challenge in identifying key QoE influence factors and developing reliable QoE models that can be used for QoE management purposes. A first research step was to conduct a systematic overview of studies addressing QoE IFs, followed by an overview of various QoE metrics and assessment methodologies. Findings were used to specify various empirical studies in both field and lab environments aimed at answering target research questions. With each study, we aimed to isolate parameter values with a final goal being to provide a straightforward approach to achieving acceptable QoE. Studies involved interactive three-party audiovisual conversations based on WebRTC technology in leisure contexts with symmetric and asymmetric device conditions. In total, six user studies (labeled as US1 to US6) employing subjective assessment were conducted over the the course of four years (Table 1.1).

The research was conducted in several phases (Figure 1.2). In the **first phase**, subjective studies were conducted to investigate end user QoE and various related features while using multiparty video conferencing services on mobile devices. The study aimed to identify the necessary topology and hardware configuration that can ensure acceptable QoE in a leisure context, together with the identification of basic QoE metrics. The questionnaire covered ratings of the

Table 1.1: Summary of conducted subjective user studies.

User Studies	Participant, MIN/MAX/AVG age	End user device	Manipulated parameters	Addressed research questions
US1, 2015, [39]	18 males, 12 females, 29/65/35	Samsung S3, S5, LG G3	Device capabilities	RQ2
US2, 2016, [40]	14 males, 13 females, 32/65/38	3 x Samsung S6	Video resolution, bitrate	RQ2
US3, 2017, [5]	16 males, 14 females, 1 fixed user per test group, 33/49/40	3 x Samsung S6	Video resolution, bitrate, frame rate, packet loss	RQ4
US4, 2018, [41]	21 males, 6 females, 20/29/21	3 x Samsung S6	Video resolution, bitrate, frame rate	RQ4
US5, 2018, [42]	7 males, 20 females, 20/25/22	3 x Samsung S7	Video resolution, bitrate, frame rate	RQ2, RQ3, RQ5
US6, 2018, [43]	16 males, 11 females, 23/23/28	3 x Samsung S7	Video resolution, bitrate, frame rate	RQ2, RQ3, RQ5

impacts of considered factors (resource and application factors) and metrics on user's QoE. In addition, used questionnaires contained questions related to user habits and opinions in a multiparty video calls. The number of participants included in the experiments was sufficient to enable deriving statistically significant results. To avoid discomfort between subjects and enable normal conversation, all participants were acquaintances. After performing the questioning, statistical analysis of obtained answers was conducted by using relevant statistical methods. Based on obtained results, key factors and QoE metrics were identified, as well as subjects' habits and opinions with respect to participating in a multiparty video call in a leisure context.

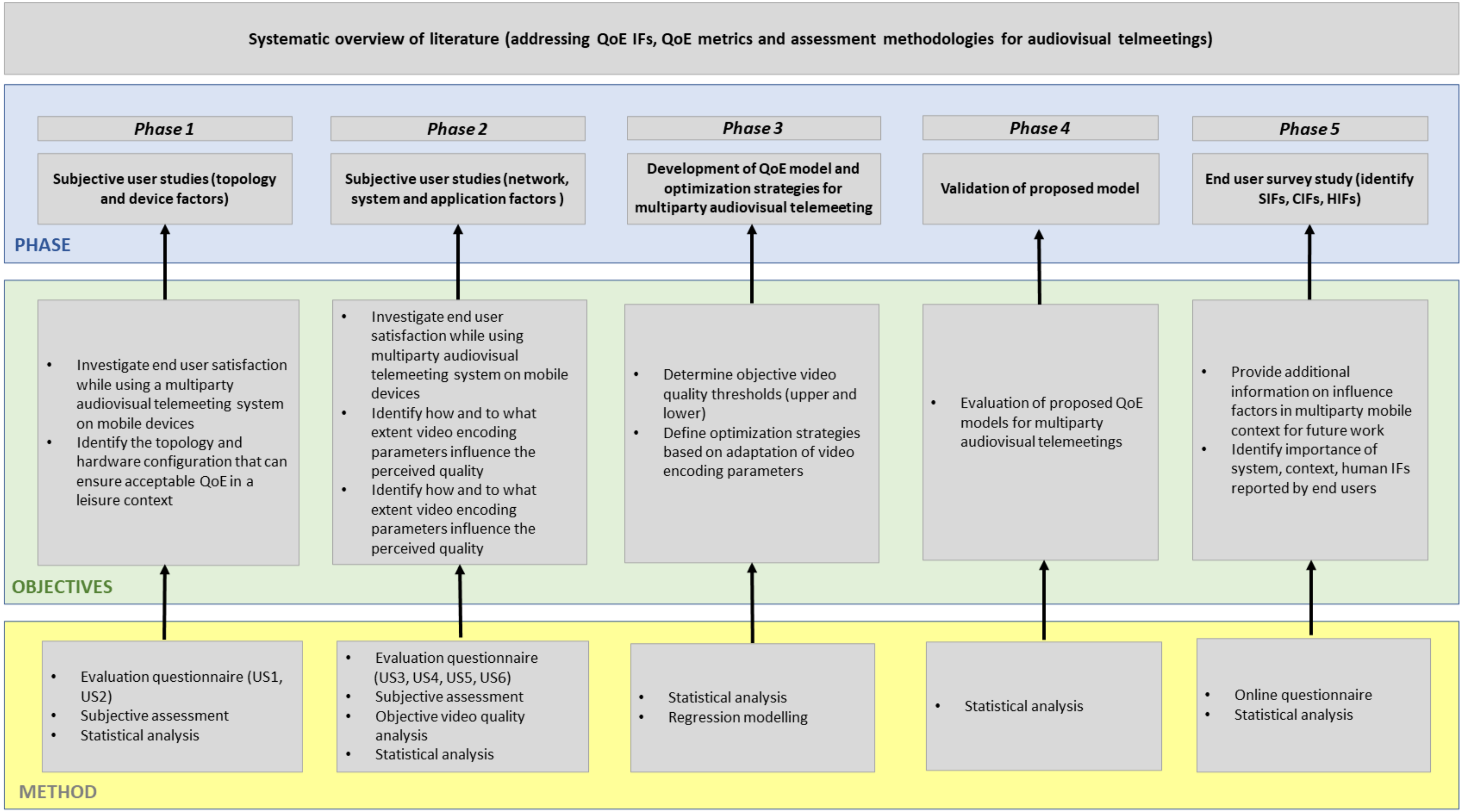
The **second research phase** included user studies with the experiments aimed to gain empirical data for the development of a QoE model for multiparty audiovisual telemeeting on mobile devices. The main goal of the user studies was to investigate how and to what extent video encoding parameters influence the perceived quality under different system and network conditions. Measurements in studies were conducted in a controlled environment, combined with subjective and objective assessment methods to study the complex relationships between network QoS parameters, application level parameters, and overall QoE. The user studies included multiparty video call sessions containing multiple test scenarios that differ according to different video encoding parameters. The corresponding combinations of the video resolution, frame rate and bitrate were tested and subjectively evaluated, without predetermined task, using free conversation which is more appropriate for audiovisual quality evaluation whereby participants will be fully focused on the display. The recommended conversation time length depends on the number of test participants, with a test time per test condition set to avoid tiredness and reduced attention of participants. Empirical data was gained by way of questionnaires and analyzed by using statistical methods. These results served as input for deriving QoE model for multiparty video calls on mobile devices.

The **third phase** of the research consisted of proposing a video encoding adaptation strategy with respect to system and network resource availability. The objective of this phase was to find thresholds (upper and lower) for video encoding parameters to yield maximum possible QoE scores while meeting resource availability constraints. The derived video encoding adaptation strategy based on the proposed QoE model is aimed to be utilized for efficient and dynamic service adaptation. In the **fourth phase**, relevant metrics were used to verify proposed QoE models. Finally, we conducted an end user survey to determine influence factors (beyond video encoding parameters) impacting multiparty audiovisual telemeetings (on mobile devices) QoE to provide the basis for enhancing and extending the model in future work.

1.4 Summary of contributions

The contributions of this thesis may be summarized as follows:

- C1: Specification of key system, contextual, and human influence factors that impact QoE and corresponding QoE dimensions in the context of multiparty audiovisual telemeetings on mobile devices.
- C2: QoE model for multiparty audiovisual telemeetings on mobile devices quantifying the relationship between objective and subjective quality metrics in a given context, based on previously established relationships between video encoding parameters and system impairments, and objective quality metrics.
- C3: QoE-driven video encoding adaptation methods for multiparty audiovisual telemeetings based on the derived QoE model.



Systematic overview of literature (addressing QoE IFs, QoE metrics and assessment methodologies for audiovisual telmeetings)

PHASE

Phase 1

Subjective user studies (topology and device factors)

Phase 2

Subjective user studies (network, system and application factors)

Phase 3

Development of QoE model and optimization strategies for multiparty audiovisual telemeeting

Phase 4

Validation of proposed model

Phase 5

End user survey study (identify SIFs, CIFs, HIFs)

OBJECTIVES

- Investigate end user satisfaction while using a multiparty audiovisual telemeeting system on mobile devices
- Identify the topology and hardware configuration that can ensure acceptable QoE in a leisure context

- Investigate end user satisfaction while using multiparty audiovisual telemeeting system on mobile devices
- Identify how and to what extent video encoding parameters influence the perceived quality
- Identify how and to what extent video encoding parameters influence the perceived quality

- Determine objective video quality thresholds (upper and lower)
- Define optimization strategies based on adaptation of video encoding parameters

- Evaluation of proposed QoE models for multiparty audiovisual telemeetings

- Provide additional information on influence factors in multiparty mobile context for future work
- Identify importance of system, context, human IFs reported by end users

METHOD

- Evaluation questionnaire (US1, US2)
- Subjective assessment
- Statistical analysis

- Evaluation questionnaire (US3, US4, US5, US6)
- Subjective assessment
- Objective video quality analysis
- Statistical analysis

- Statistical analysis
- Regression modelling

- Statistical analysis

- Online questionnaire
- Statistical analysis

Figure 1.2: Research methodology per research phase.

1.5 Thesis structure

The thesis is structured as follows: after the introductory chapter, Chapter 2 provides an overview of multiparty audiovisual telemeeting architectures and basic concepts. Chapter 3 gives a state-of-the-art literature review and overview of relevant standards in the field of multiparty telemeeting QoE. Chapter 4 contains the results of an end user survey involving 272 participants conducted to obtain insights into user's habits and opinions with respect to multiparty telemeetings and QoE expectations. Chapter 5 reports on the results of two initial user studies (US1 and US2) that investigate the impact of end user device capabilities and different video encoding parameters on end user's QoE. Furthermore, in Chapter 5 we also report on the results of user studies (US5 and US6) addressing the impact of system factors. Additionally, we investigate the relationship between objective video quality metrics (blurriness and blockiness) and subjective quality ratings. We empirically derived estimation models for QoE and perceived video quality. The QoE model quantifies the relationship between QoE and perceived video and audio quality, while the perceived video quality model quantifies the relationship between objective (in terms of video encoding parameters, and in terms of blurriness and blockiness) and subjective quality in a given context. Subsequently, Chapter 5 includes the validation process and the accuracy of the proposed QoE models. In Chapter 6 we describe user studies (US3 and US4) addressing the impact of network influence factors on the perceived video quality and the QoE. Based on the obtained results from conducted subjective studies, we propose in Chapter 7 novel QoE-driven video adaptation strategies for three-party telemeetings established via smartphone devices in leisure contexts. Finally, Chapter 8 concludes the thesis, summarizes the contributions, discusses the limitations, and provides an outlook for future work.

Chapter 2

Multiparty audiovisual telemeetings

The rise of multiparty audiovisual telemeetings in response to the pandemic comes with the challenges in terms of Internet usage and traffic volume. Aiming to provide acceptable QoE, it is necessary to understand underlying technologies. This chapter provides an overview of multiparty telemeeting architectures in Section 2.1, and then focuses on WebRTC technology, its topology, protocols and media processing. Finally, in Section 2.2 we give an overview of existing audiovisual telemeeting platforms.

2.1 Multiparty audiovisual telemeeting architectures

Standards organizations responsible for specifying relevant architectural components, protocols, and quality assessment methods and models for audiovisual telemeetings include the International Telecommunications Union (ITU) and the Internet Engineering Task Force (IETF). While the IETF standardizes protocols such as Real-Time Protocol (RTP) and Session Initiation Protocol (SIP), ITU-T video conferencing standards cover the higher service level. Additionally, the Video Quality Expert Group (VQEG) brings together experts in subjective video quality assessment and objective quality measurement to combine individual research efforts into general solutions, thus connecting experts from industry, academia, government organizations, ITU, and other Standard Developing Organizations (SDO), with the goal to enhance the field of video quality of television and multimedia applications [44]. VQEG conducts subjective video quality experiments, validates the objective video quality models, and designs new methods.

ITU-T recommendation P.1301 on “Subjective quality evaluation of audio and audiovisual multiparty telemeetings” defines the terms telemeeting and multiparty [17]. A **telemeeting** is defined as a meeting in which participants are located in at least two different locations and the communication takes place via a telecommunication system. **Multiparty** refers to service scenarios involving more than two participants located at two or more locations.

Telemeetings can be used in conventional business video conference scenarios, as well as

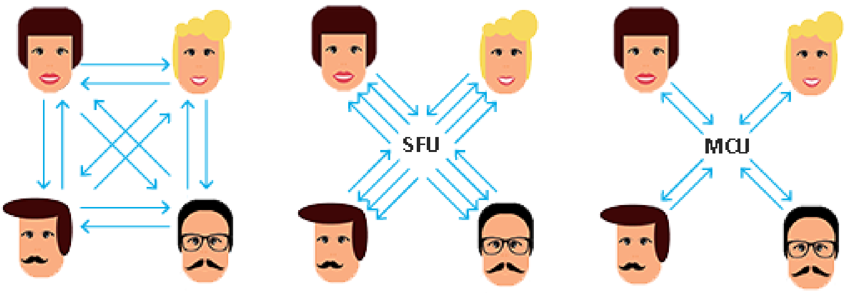


Figure 2.1: Different topologies used during multiparty calls: Peer-to-Peer, Selective Forwarding Unit (SFU), and Multipoint Control Unit (MCU).

more flexible private meetings in a leisure context [20]. In general, telemeetings organized in a business context have specific objectives and agendas, with a set of tasks that must be completed, while telemeetings held in a private/leisure context have the primary objective of experiencing a sense of presence or social connection [17]. Different objectives correspond to the different meeting contexts, whereas expectations in the leisure context might be lowered [36], [6]. We note that traditionally the term *video conferencing* was related to the business context, while the term *video call* was more commonly associated with the leisure or private context. We therefore consider both aforementioned terms as referring to the more general term *audiovisual telemeetings*, and use these terms interchangeably.

With respect to technical realization, several connection types are possible for multiparty audiovisual telemeetings, as illustrated in Figure 2.1. One possibility is a **full-mesh topology**, where in communication with n peers, each peer handles $n-1$ download streams and $n-1$ upload streams (for illustration purposes, the figure portrays each peer transmitting one stream). Peer-to-Peer topology is most affordable but requires a high amount of processing power, lacking in older smartphones, and higher capacity in terms of available bandwidth. To release the load on both the end user device resources as well as the network, part of the processing and data transmission burden may be shifted to a centralized media server, albeit with potentially higher operational costs (due to administration, signaling, and media distribution) [45]. A **Selective Forwarding Unit (SFU)** requires peers to upload their own stream, and distributes it to all other connected peers. Each peer handles one upload stream and $n-1$ download streams (for illustration purposes, we assume each peer to be transmitting one stream). Finally, peers connected to a so-called **Multipoint Control Unit (MCU)** generally handle one upload and one download stream, while the MCU is responsible for mixing uploaded streams into a single stream, adapting streams, and distributing to other peers [46].

Until recently, video conferencing was primarily used by larger organizations who could afford the endpoints (conference rooms), infrastructure and personnel required to implement and maintain an on-premise solution. Eventually, endpoints became widespread desktop computers, but still used mostly within larger organizations. As the number of endpoints increased, the needed infrastructure become more complex, resulting in higher operational costs. Operational overhead can be reduced with cloud-based solutions, whereas service providers ensure the necessary infrastructure. Most conferencing services nowadays utilize various cloud-based solutions which enable scalable service architectures and support for high quality audiovisual communication. Moreover, emerging technologies such as WebRTC offer free and open source solutions for building real-time communication services. Given the widespread and increasing use of WebRTC, both in commercial and free video communication platforms and services (e.g., Google Meet, Microsoft Teams, Big Blue Button, Whereby, Talky, FaceTime), we have used WebRTC technology in the context of the studies reported in this thesis. We thus give a brief overview of the WebRTC architecture, APIs, protocols, and codecs in the following section.

2.1.1 WebRTC architecture and protocols

WebRTC is an open project that provides APIs for developing applications to capture and stream real-time audio and video, as well as to exchange data between browsers without requiring any plug-ins or any other third-party software [47]. Mobile devices can support WebRTC via custom built applications. An application can then be developed as a hybrid application relying on third-party software, or a native app built for a specific mobile operating system such as Android or iOS.

WebRTC communication is encrypted, where sessions can be established directly between browsers, providing security and privacy. WebRTC is part of the HTML5 proposal and enables web applications to use RTC functionalities using JavaScript APIs, providing video, audio, and data exchange. The HTML format is used in the scope of web applications, and JavaScript enables users dynamic interaction with the web application. The APIs can be defined as a set of methods that enable developers to access the available technologies supported by browsers and are used to create web pages and interactive web applications. The basic WebRTC architecture includes a server and at least two participants (Figure 2.2) [46]. A common use case involves an application being downloaded from the same web site in a local environment (browser), while a communication server is required to establish communication between the browsers. Applications use the WebRTC API to communicate with the local context.

In general, a user initiating a video call creates an online virtual room via a web application to start a WebRTC session. Other users are invited to access the room through their web browser and generated URL. Such communication requires negotiation of the media path between caller and callee, and the APIs offering the JavaScript application access to the browser's functionalities [46]. WebRTC uses UDP (User Datagram Protocol) as a transport protocol for real-time communication. However, sometimes due to the network address translators (NAT) or firewalls, peers are not able to connect directly, hence additional mechanisms are needed to traverse the NATs and firewalls such as ICE (Interactive Connectivity Establishment), STUN (Session

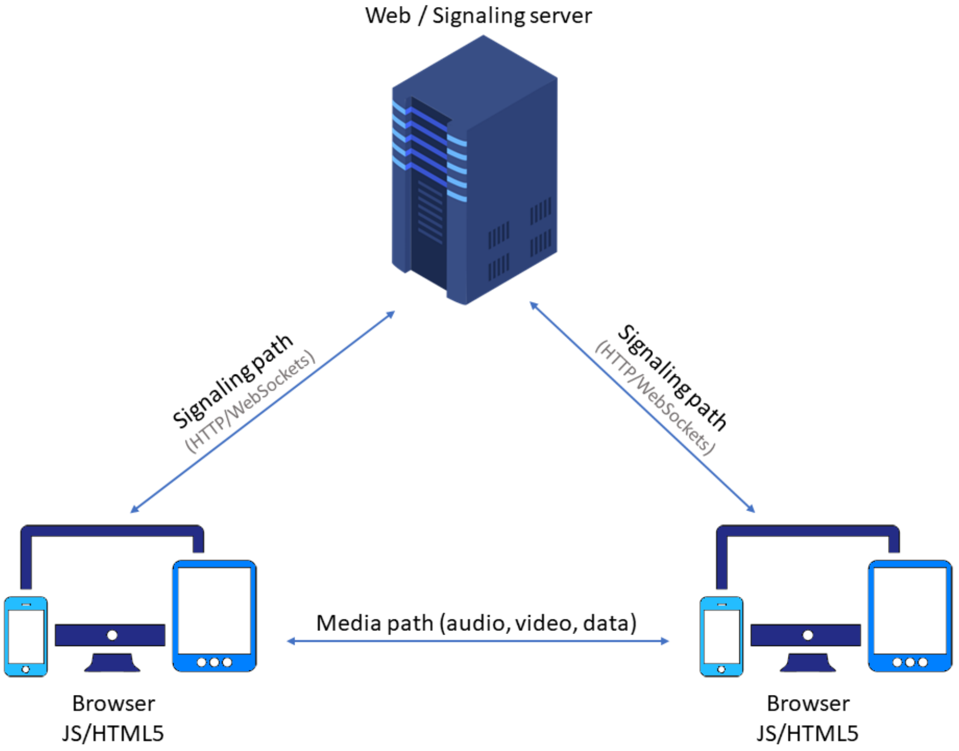


Figure 2.2: Basic WebRTC architecture.

Traversal Utilities for NAT), and TURN (Traversal Using Relays around NAT) [48], [49], [50]. ICE allows the endpoints to detect the type of NAT using STUN and TURN protocols to find a path for connection establishment. A STUN server is used to obtain an external network address, while the TURN server is used to relay traffic if direct P2P session establishment is not possible. Encryption is a mandatory part of WebRTC and is enforced with DTLS (Datagram Transport Layer Security). DTLS exchanges the keys that SRTP (Secure Real-time Protocol) uses for the encryption. When keys used to encrypt the stream at both peers are exchanged, the browser can start with media streaming over SRTP (Figure 2.3).

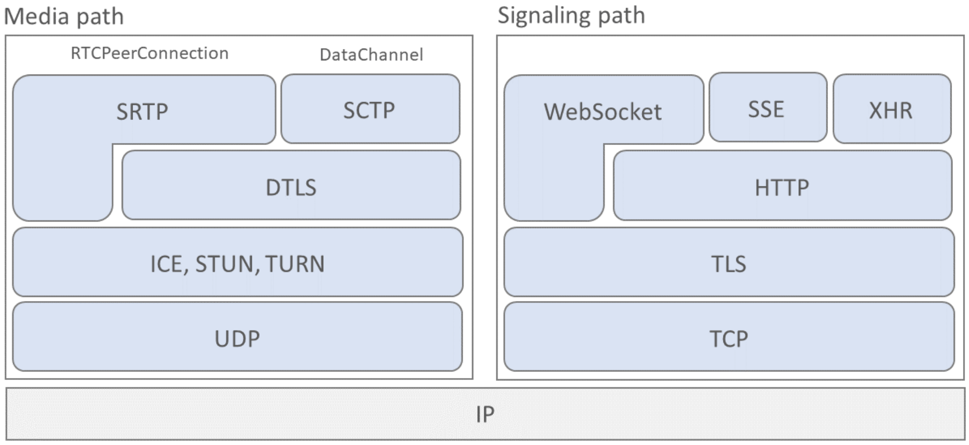


Figure 2.3: WebRTC protocol stack [51].

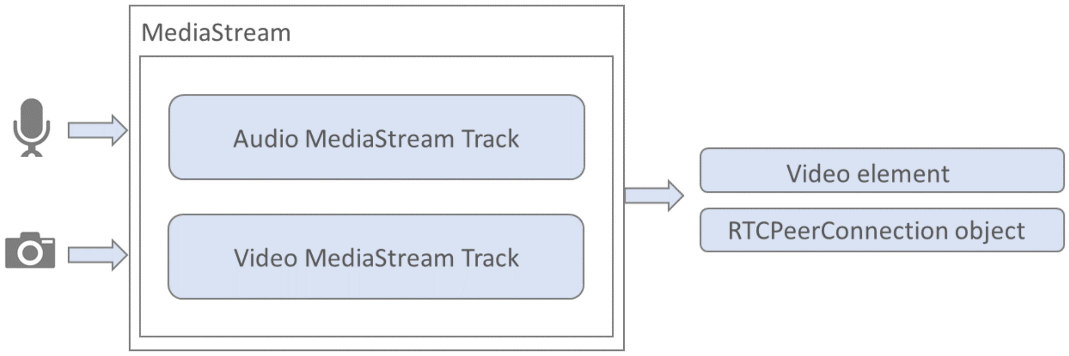


Figure 2.4: A MediaStream object combines input and output of all the object's tracks [52].

WebSockets and HTTP are available to the JS application through the browser API. When the session from the caller is initiated to the callee browser, messages are transferred via a signaling protocol (carrying SDP messages) to the web server. Session description specifies the information needed for the establishment of the media path: transport and Interactive Connectivity Establishment (ICE) information, along with other types of capabilities such as media type and media format. Upon received messages, the callee JS replies.

WebRTC supports three main APIs: MediaStream, PeerConnection and DataChannel. The **MediaStream API** is designed to handle actual audio/video streams of data. The API manages actions on the media stream such as: access to the media streams from local cameras and microphones, display of the stream's content, or sending of the stream to a remote peer. Security is ensured through user permissions - before joining the room, each user must grant the service to access their camera and microphone. A WebRTC application requests access to local media devices with the getUserMedia() function. As a result, a media stream is returned to the application that can be played in the HTML5 video element or can be sent as output to the RTCPeerConnection object, which then sends it to a remote peer. *GetUserMedia* returns a *MediaStream* object which contains MediaStreamTracks that transfer the actual video and audio stream, (Figure 2.4) [46]. The MediaStreamTrack object may include several channels (right and left audio channels).

The **PeerConnection** API involves all the internal mechanisms of the browser that enable media and data transfer between peers. It also handles with JavaScript methods the exchange of signaling messages. WebRTC enables two ways to establish a session: one is using SIP WebSocket (WebRTC SIP/WS), and the other by JSON (JavaScript Object Notation) XMLHttpRequest (WebRTC JSON/XHR) requests. Both approaches are based on JSEP (Javascript Session Establishment Protocol), similar to the SIP offer/answer signaling model [53]. The PeerConnection interface represents a connection between the local browser and a remote peer. To ensure direct connection between browsers, PeerConnection uses the ICE protocol together with the STUN and TURN servers.

The **DataChannel** API, based on the RTCDataChannel and leveraged through RTCPeer-

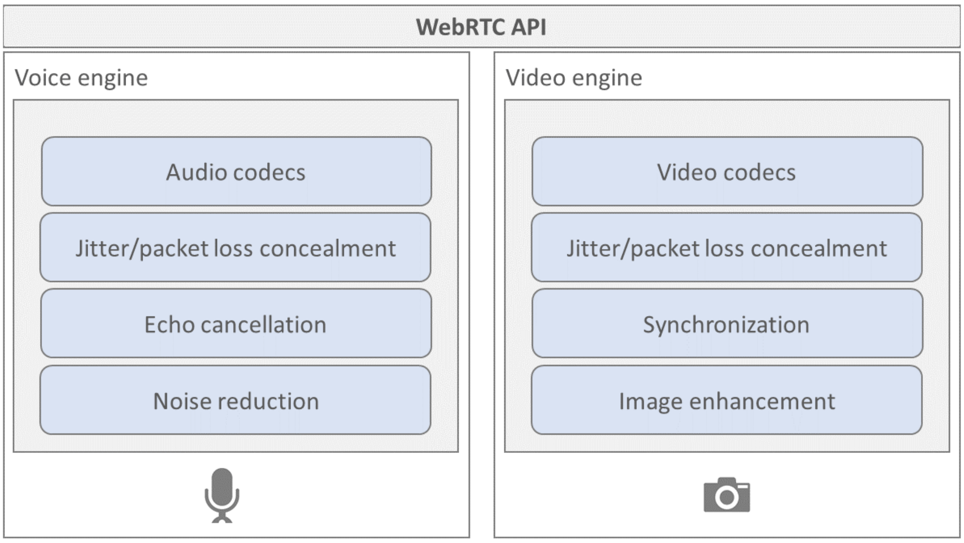


Figure 2.5: WebRTC audio and video engine.

Connection, enables bidirectional data exchange between users, where each stream represents an unidirectional logical channel. Each channel must support in-order or out-of-order delivery as well as reliable or unreliable delivery. The data channel can be secured through the encapsulation of SCTP (Stream Control Transmission Protocol) over DTLS and UDP, and together with ICE mechanism enables confidentiality, source authentication, and integrity protected transfers [46]. SCTP is used as an application protocol since it multiplexes multiple independent channels, provides congestion and flow control, and provides reliable or partially reliable delivery.

2.1.2 Media processing and coding in WebRTC

Raw media captured by the end user device is sent to the browser for signal processing. This is done directly by the browser based on the voice engine and video engine framework. Both engines include specific media codecs, for voice (e.g., Opus, G.711, iSAC, and iLBC) and video (e.g., VP8, VP9, H.264) and methods for error concealment to mask the disturbances of jitter and packet loss, (Figure 2.5). The video engine is responsible for synchronization and image enhancements, while the audio engine applies echo cancellation algorithms and noise reduction to enhance the audio quality.

Raw streams are processed to enhance quality adjusting the media quality to the available network resources such as bandwidth, jitter, and latency delays. The processed signal is then passed to the encoder for data compression. Audio samples and video frames are sent for packetization (Figure 2.6). After the packetizer builds packets with application specific headers, packets are ready to be sent to the queue, and the sender buffer is responsible for scheduling a packet for transmission. On the receiving side, packets are transferred to the audio and video

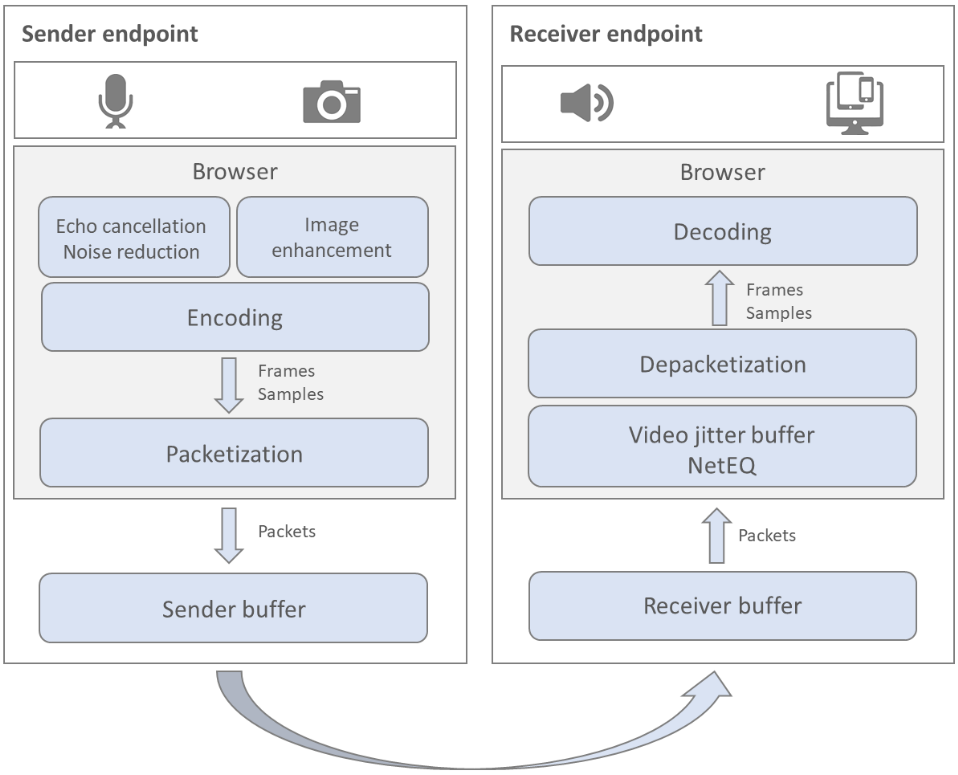


Figure 2.6: WebRTC media processing pipeline.

jitter buffer (implemented in the browser), a temporary storage buffer used to order, synchronize or drop late packets. When packets arrive for the de-packetization, headers are removed and packets are split to the samples and frames which are forwarded to the decoding process for interpretation and translation to the raw data ready for display.

Codecs

Mandatory implemented audio codecs in the scope of the WebRTC standard include Opus and G.711. Opus is defined by RFC 6716 and is typically used in WebRTC sessions [54]. Opus supports bitrates from 6 kbit/s to 510 kbit/s. Given a 20 ms frame size, the following bit rates are recommended for different configurations:

- 8-12 kbit/s for narrow-band speech,
- 16-20 kbit/s for wide-band speech,
- 28-40 kbit/s for full-band speech.

With respect to video, endpoints and corresponding web browsers must implement H.264 and VP8, while the application can choose which one will be used. Web browsers are actively adding VP9 support as well, even though it is an optional video codec in WebRTC. The VP8 royalty free codec is often used in WebRTC applications. VP8 uses a lossy DCT-based algorithm, and allows arbitrary bitrates as well as frame rates. Supported frame sizes are up to 16384 x 16384 pixels and mode YUV 420 color sampling at 8 bits per channel depth. VP8 is based on two frame types: intraframes and interframes [55]. Intraframes, known as a key frames

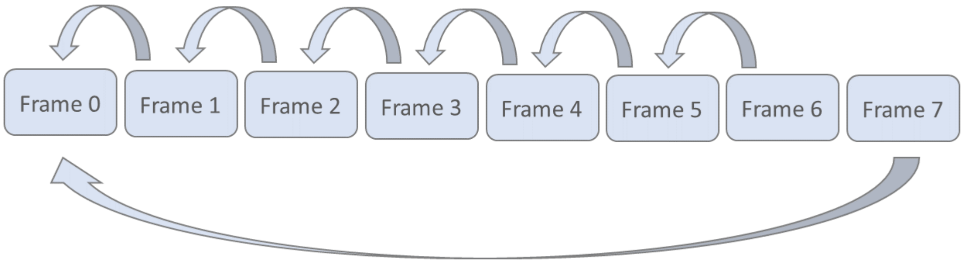


Figure 2.7: VP8 error recovery.

are decoded without reference to any other frame in a sequence. Interframes are encoded with reference to prior frames, specifically all prior frames up to and including the most recent key frame. If frame 0 is a key frame, subsequent frames from one to six build predictors from only the prior frame, while the seventh frame only uses frame 0 as a reference, and as such can still be decoded even if the previous frames from one to six are lost (Figure 2.7).

Congestion control

With video conferencing/telemeeting services typically designed to use UDP rather than TCP (Transmission Control Protocol), the deployment of congestion control mechanisms is left to the application layer [18]. As such, the Google Congestion Control (GCC) algorithm has been specifically designed to work with RTP/RTCP protocols and target real-time streams such as telephony and video conferencing. The goals for any congestion control algorithm are:

- preventing network collapse due to congestion,
- allowing multiple flows to share the network fairly.

The GCC algorithm was proposed within the RTP Media Congestion Avoidance Techniques (RMCAT), an IETF Working Group that aims to develop new protocols which can manage network congestion in the context of RTP streaming [56]. The GCC algorithm includes two control elements: a delay-based controller on the receiver side, and a loss-based controller on the sender side (which complements the delay-based controller if losses are detected). The congestion controller bases decisions on how to invokes stream adaptation, including bitrate, resolution, and frame rate, on measured round-trip time, packet loss, and available bandwidth estimates.

Additional algorithms developed for interactive real-time media applications (such as video conferencing) including congestion control are Network-Assisted Dynamic Adaptation (NADA) [57], developed by Cisco, and Self-Clocked Rate Adaptation for Multimedia (SCReAM) [58], developed by Ericsson. SCReAM was originally designed for WebRTC, but it can be used in other applications where RTP congestion control is needed. Both algorithms adapt sending rate based on the delay and packet loss control mechanisms.

2.2 Overview of platforms for audiovisual telemeetings

In October 2016, mobile and tablet Internet usage was reported to exceed desktop usage for the first time worldwide [59]. Nowadays, and especially in light of the ongoing COVID-19 pandemic, we are witnessing a drastic increase in the use of video conferencing tools, for purposes such as e-learning, meetings, and social gatherings [60]. When face to face communication is not an option, video conferencing can offer a remote and flexible work environment and the feeling of social presence. Video conferencing has become a tool for enabling diverse applications and communication scenarios, ranging from socializing with family and friends, to offering remote healthcare-related tele-consultation services, providing psychotherapy to those in need, to various business related meetings and events [61], [62], [63].

The platforms need to provide a wide range of features and tools that can help to conduct quality collaboration. For interactive meetings, content sharing is often a key feature. Sharing of screen, content, or applications makes it easier for users to follow and contribute to the meeting. Another important feature is that the face of the active speaker can be seen during the meeting in order to pick up on important visual cues. For additional interactions, support for texting during the meeting can be very useful. For collaborative meetings, remote control which offers the possibility to make changes on the shared content in real time is also a valuable feature. To store everything of importance, especially in a working environment, the host and/or participants should be able to record the meeting, save it, and afterwards easily share it. Finally, tracking of various performance and utilization metrics may be of importance, including factors such as number of participants, concurrent connections, or geographic distribution.

Embracing video conferencing technology, especially in 2020, can rescue businesses and even prevent loneliness in a time of social distancing. There are numerous applications and platforms available on the market for video telemeetings. While video conferencing was previously both expensive and complex, nowadays available video conferencing services are easy to manage and affordable or even free. Naturally, industry leaders such as Cisco Webex solutions can be costly, but at the same time on the other end of the price spectrum web applications with tailored plans to suit different needs are gaining popularity. Among various popular communication tools which have seen a high increase in customer use, such as Zoom and Microsoft Teams, included are also numerous applications based on WebRTC technology (e.g., Whereby, Uberconference, Google Meet, BlueJeans, Lifesize, Slack) (Table 2.1) [64]. All applications listed in the table use encryption and provide additional functionalities such as texting, content sharing and recording.

Table 2.1: Video conference services.

Service	Number of participants: free / paid subscription	Linux/MacOS/Windows	Mobile device support	Cloud-based	Video quality	License
Cisco Webex [65]	N/A / 1000	x / ✓/ ✓	✓	✓	VGA, HQ, HD	Proprietary
Adobe Connect [66]	N/A / 100	Partial / ✓/ ✓	✓	✓	VGA, HQ, HD	Proprietary
Fuze meeting [67]	N/A / 1000	Partial / ✓/ ✓	✓	✓	QVGA, HD	Proprietary
Intermedia AnyMeeting [68]	200	x / ✓/ ✓	✓	✓	HQ	Proprietary
Google Meet [69]	100 / 250	✓/ ✓/ ✓	✓	✓	HD	Proprietary
Jami [70]	Unknown	x / ✓/ ✓	✓	✓	VGA, HQ, HD	GNU General Public License
Jitsi Meet [71]	50 / N/A	✓/ ✓/ ✓	✓	x	VGA, HQ, HD	Apache License
Lifesize [72]	10 / 300	x / ✓/ ✓	✓	✓	HQ	Proprietary
GoToMeeting [73]	N/A / 250	✓/ ✓/ ✓	✓	✓	VGA, HD	Proprietary
Microsoft Teams [74]	300 / 10000	✓/ ✓/ ✓	✓	Partial	VGA, HQ, HD	Proprietary
Skype [75]	100 / N/A	✓/ ✓/ ✓	✓	x	VGA, HQ, HD	Proprietary
Skype for Business [76]	N/A / 1000	x / ✓/ ✓	x	Partial	VGA, HQ, HD	Proprietary
StarLeaf [77]	20 / 100	✓/ ✓/ ✓	✓	✓	HD	Proprietary
TrueConf [78]	12 / 1600	✓/ ✓/ ✓	✓	✓	Ultra HD	Proprietary
VideoMost [79]	N/A / 300	✓/ ✓/ ✓	✓	✓	HD, FHD, Ultra HD	Proprietary
WizIQ [80]	1999	✓/ ✓/ ✓	x	Partial	VGA, HQ, HD	Proprietary
Zoom [81]	100 / 1000	✓/ ✓/ ✓	✓	✓	VGA, HQ, HD	Proprietary
Blue Jeans [82]	N/A / 100	✓/ ✓/ ✓	✓	✓	HD	Proprietary
Whereby [83]	4 / 50	✓/ ✓/ ✓	✓	✓	HD	Proprietary
Uberconference [84]	10 / 100	✓/ ✓/ ✓	✓	✓	HD	Proprietary

2.3 Chapter summary

This chapter gives an overview of multiparty telemeeting communication topologies and architecture design, focusing in particular on the WebRTC paradigm which was utilized across the different studies conducted in the scope of this thesis. Design and specific implementation of a audiovisual telemeeting service depend on the context, space, usage scenario, size of a meeting, and users' needs. Designers need to take into consideration the hardware possibilities of target endpoints, and factors such as audio (microphone and speaker setup), and video (single/multiple screens, size of the screen, and camera quality) requirements as well as available resources of the communication channel. Thus, when requesting audio and video, special attention should be paid to the quality of the streams. While the hardware may be capable of capturing high quality streams, the CPU and bandwidth capacity must be sufficient to smoothly process multiple receiving streams.

Different platforms offer various features to enhance the meeting quality in terms of collaboration, engagement, and interaction, ultimately making the communication easier and more effective. However, to enhance overall user experience, it is necessary to understand the role of key underlying system-, context-, and human-related influence factors. Thus, the following chapter introduces basic QoE concepts in light of audiovisual telemeetings and related influence factors.

Chapter 3

State-of-the-art review: QoE for multiparty audiovisual telemeetings

Following the overview of multiparty audiovisual telemeeting characteristics given in the previous chapter, this chapter first introduces general definitions and assessment methods related to QoE for video conferencing services in Section 3.1, followed by an overview of the challenges related to assessing QoE for multiparty audiovisual telemeetings in Section 3.2. In Section 3.3 we give an overview of studies addressing specifically QoE for multiparty audiovisual telemeetings. Finally, the chapter is concluded with a summary in Section 3.4.

3.1 Quality of Experience definitions

Over the years, different bodies aimed to define QoE with several transitional forms evolved. According to ITU-T Recommendation P.10/G.100 from 2007, QoE is defined as the *"overall acceptability of an application or service, as perceived subjectively by the end user"* [85], where QoE includes the complete end-to-end system effects and may be influenced by user expectations and context. The European Telecommunications Standards Institute (ETSI) defined QoE as *"a measure of user performance based on both objective and subjective psychological measures of using an ICT service or product"* [86]. Objective psychological measures do not rely on the user opinion (e.g., time to complete task, task accuracy measured in number of errors), whereas subjective psychological measures do rely on the user opinion (e.g., perceived quality of medium, satisfaction with a service). Thus, whenever we think about QoE we have to think beyond objective metrics or measures. The main goal of QoE is to capture perceived quality. The term *perceived* brings subjective humans into the scope and colors it with diversity. In social psychology, human perception refers to the different mental processes that we use to form impressions and conclusions towards presented stimuli [87]. Humans make judgments and opinions in accordance with particular emotions, based on previous experience and expectations they might have related to a given subject.

To establish definitions and methods for the quantitative assessment of QoE for multimedia content and services in a given situation and system configuration, one of the goals of the COST IC 1003 Action - Qualinet was to clarify the definition of Quality of Experience based on terms "Quality" and "Experience" [88]:

- ***Experience**: "An experience is an individual's stream of perception and interpretation of one or multiple events".*
- ***Quality**: "The outcome of an individual's comparison and judgment process. It includes perception, reflection about the perception, and the description of the outcome".*

As a result one of the most commonly cited definitions of QoE was defined as *"the degree of delight or annoyance of the user of an application or service. It results from the fulfillment of his or her expectations with respect to the utility and/ or enjoyment of the application or service in the light of the user's personality and current state"* [88].

The multidimensional nature of QoE stems from a number of different influence factors and perceived features comprising the overall QoE. The Qualinet white paper further defines an **influence factor** 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"* and groups IFs into three categories [88] [89] [90]:

- ***Human IF (HIF)** as any variant or invariant property or characteristic of a human user. The characteristic can describe the demographic and socio-economic background, the physical and mental constitution, or the user's emotional state.*
- ***System IFs (SIFs)** referring to the properties and characteristics that determine the technically produced quality of an application or service. They are related to media capture, coding, transmission, storage, rendering, and reproduction/display, as well as to the communication of information itself from content production to user.*
- ***Context IFs (CIFs)** factors that embrace any situational property to describe the user's environment in terms of physical, temporal, social, economic, task, and technical characteristics.*

The complexity in assessing and modeling QoE arises not only from the multitude of different system, context, and user IFs [91], but from the difficulty in controlling certain factors during studies. As discovered during conducted studies, certain factors (e.g., video quality was lowered due to the CPU overuse), may be unintentionally manipulated during the course of tests, hence impacting user ratings (as illustrated in Figure 3.1). Furthermore, quality assessment is complicated given the multidimensional nature of QoE itself, in terms of perceived quality dimensions.

In the following sections, further details are given with respect to specific IFs and QoE dimensions (referred to as *QoE features*) relevant for assessing the QoE of multiparty audiovisual

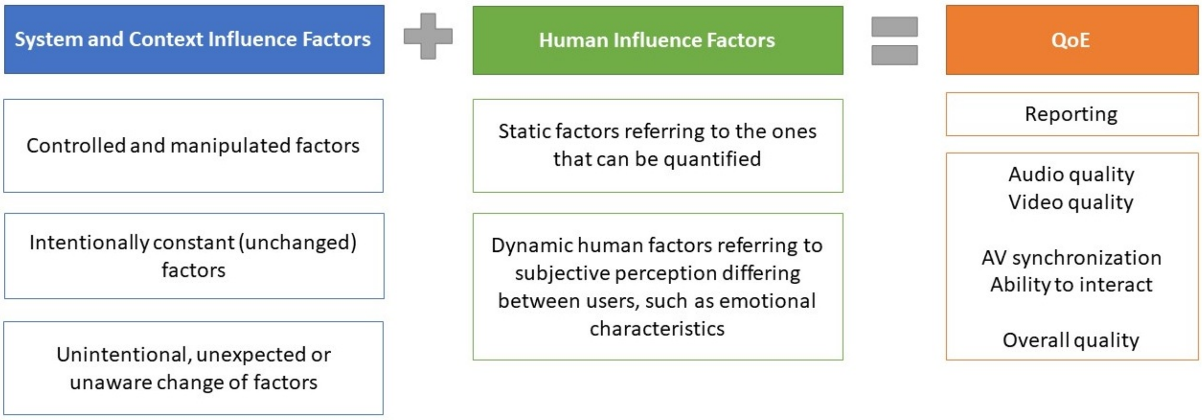

System and Context Influence Factors
Controlled and manipulated factors
Intentionally constant (unchanged) factors
Unintentional, unexpected or unaware change of factors
Human Influence Factors
Static factors referring to the ones that can be quantified
Dynamic human factors referring to subjective perception differing between users, such as emotional characteristics
QoE
Reporting
Audio quality
Video quality
AV synchronization
Ability to interact
Overall quality

Figure 3.1: Factors influencing QoE.

telemeetings in mobile environments, with the aim being to identify and further explore those that are identified to be the most influential.

3.1.1 QoE influence factors

Modern mobile users want to be able to access and reliably utilize demanding services regardless of context or system influence factors, such as location, time, network conditions, service topology, or mobile device processing capabilities. Mobile end user devices such as smartphones used to take part in multiparty video calls often represent possible bottlenecks in the service delivery chain. Fortunately, each new generation of device has brought more advanced hardware in terms of memory, processor power, camera, and battery cycle. The smartphone display size tended to get larger as well, trying to accommodate higher resolution screens. The resolution 1080x1920 pixels was the screen resolution implemented in high-end smartphones in 2014. Five years later it become the most common resolution [92]. Rapid development in the smartphone hardware industry in the last several years implies that majority of recently released smartphones should be able to provide acceptable QoE for multiparty video calls with adapted video quality streams.

In the context of mobile networks, characterized by variable resource availability, challenges arise with respect to meeting the QoE requirements of conversational real-time, media rich, and multiparty services [26]. In addition to network requirements, multiparty video conferencing services impose high requirements on end user device processing capabilities, with the need for real-time encoding and decoding of multiple media streams.

Skowronek *et al.* identified mobility, device and encoding interoperability, ease of use, and additional collaboration possibilities (e.g., exchanging pictures, files, chatting) as the most important aspects for telemeeting services [21]. However, mobility and device capacity, alone and together, can create asymmetry, which is a realistic and common case. Thus, the impact of the mobility and device, depending on the number of the participants, can greatly differ due to the numerous possible combinations of connection type between locations and type of equipment being used.

Based on standards, related work, and own research, a detailed overview of QoE IFs for multiparty telemeetings in a mobile context according to identified categories is given, as summarized in Table 3.1.

System influence factors

Acceptable end-to-end delay and jitter values for audio and video streams (discussed in the following paragraphs) in general do not significantly impact QoE if media buffers are available. While conversational speech is highly sensitive to delay, jitter, and loss, video can also be very sensitive to delays and losses, resulting in a negative impact on the perceived quality of interactive services such as audiovisual telemeetings. Depending on the information being lost, produced disturbances introduce different impairments. Thus, impairments and their impact level depend on the type of data lost (system information, header losses or packet types I, P, B), used codec, and decoder packet loss concealment techniques [93].

Packet loss and delay: Experiments conducted over Wi-Fi and cellular network, based on the two-party video call (established between smartphone and computer with injected video sequence on both sides), showed sensitivity to packet losses and long packet delays [23]. De Moor *et al.* in the experiment involving two-party WebRTC-based audiovisual call evaluated the impact of impaired video (with a 20% packet loss), impaired audio (with restricted CPU usage on the client WebRTC application) and both streams, audio and video with 500 ms delay and 300 ms jitter [30]. Results showed that disturbances in both audio and video had the most negative impact on overall quality, while video-only impaired scenarios performed somewhat better than audio-only impaired scenarios.

Face-to-face interaction includes both body signals and vocal patterns. Spontaneous conversation is based on alternating turn talking, usually with interruptions and double talk. Transition between turns is very fast, typically 200 ms, meaning that greater extension of this period will be noticed by the participant. According to ITU-T Rec. G.114, acceptable mouth-to-ear delay is considered to be under 300 ms [94], establishing delay as a significant influence factor when it comes to interactivity. Xu *et al.* [28] investigated how to increase mouth-to-ear delay within just noticeable differences, but without perceptible reduction on interactivity. The authors suggested adding a 100 ms buffer for jittery flows. In [27] authors identified that in multiparty conversations, one-way delay (between 500 ms and 1000 ms) is perceived as a longer pause and causes double talk. Due to the delay, it is not possible for participants to just start talking when they want to be active. The structure of the conversation has to change, meaning that when an inactive participant wishes to speak, the active speaker should hand the turn verbally to him/her.

QoE influence factors related to the direct perception of speech quality include *listening*

Table 3.1: Influence factors to be considered when assessing and modeling multiparty audiovisual telemeeting QoE.

Category	Influence factors
System	
Network	Wireless channel characteristics (noise, fading, interference, interception, security)
	Wireless capacity (speed, coverage, bandwidth)
	Signal strength
	Transmission impairments: packet loss, delay, jitter, duration of impairment
Device	Processing capability, storage (temporary and permanent) capacity
	Power consumption, battery lifetime
	Video capturing: camera quality
	Placement of camera in relation to the telemeeting participant: viewing distance, angle
	Video preview: display size, resolution, number of displays
	Audio capturing: microphone quality and placement
	Audio listening: number of, type of, and placement of loudspeakers
	Endpoint with headset
	Type and version: operating system, browser - interoperability
Application	Loading time, content type (e.g., text, images, audio, video)
	Connection time when establishing a call
	Screen layout
	Current talker marking/switching
	Additional functionality (link , data and screen sharing, chat, mute, hold, record)
	Price
	Scalability, availability, reliability, security
	Ease of use
Context	
	Mobility (still, pedestrian, vehicle)
	Time of day
	Business or private use-case
	Laboratory or natural environment
	Number of participants
	Site distribution: one or several participants at two or several sites
	Room acoustics, room size
	Background noise characteristics, reverberation, illumination
	Background and clothes colors
Type of task	Task with specific goal
	Natural, spontaneous conversation with interruptions and double talk
Human	
	Personality, gender, age, pitch and level range, voice timbre
	Language
	Hearing and viewing abilities
	Emotional characteristics: amused, amazed, annoyed, bitter, bored, calm, comfortable, confident, delighted, depressed, frustrated, strong, sympathetic
	Intellectual ability, previous experience and expectations, level of education,
	Specific knowledge on certain topic
	Acquaintance of the participants

features (intelligibility, fidelity, coloration, noisiness, discontinuity, loudness), *talking features* (echo, reduced double-talk, non-optimum sidetone), and *conversational features* [95]. While the difficulty in understanding other participants can impact natural flow, restricted talking capability can impact conversational quality. Based on a conducted survey involving 140 participants, Husić *et al.* identified the following seven factors as having the strongest impact on user satisfaction in the case of WebRTC video calls: audio quality, image quality, quality of service, service price, loss of video frames, ease of use, and procedure of accessing web environment [96]. Based on this classification, García *et al.* proposed the following key performance indicators for QoE estimation: call establishment time, end-to-end delay, perceived audio, video, and audiovisual quality [97].

Video quality: Studies exploring video quality for telemeeting scenarios in different contexts with combinations of factors such as resolution, encoding bit rate, viewing distance, and up-scaling of video formats found that bit rate and viewing distance were the most significant factors affecting subjective video quality [98]. With respect to various applications and devices, studies have shown that video quality also depends on display size, resolution, brightness, contrast, sharpness, colorfulness, and naturalness [99] [100]. Zinner *et al.* conducted a subjective study showing video clips with different video quality levels to subjects. Authors investigated the influence of video quality in terms of video resolution, video frame rate, and video content types on QoE. Results showed that bandwidth saving should be accomplished by decreasing the resolution. Consequently, studies found that it is better (from a QoE perspective) to reduce resolution rather than frame rate, when faced with limited bandwidth availability [31]. The same rule should be applied in mobile environments with interactive services with low to medium motion, such as video call.

The efficiency of video compression may be considered in terms of achievable compression ratio with minimal or non-perceivable quality degradation. High compression ratios lead to perceptual spatial or temporal artifacts. Spatial artifacts such as blocking, blurring, ringing, basis pattern effect, and color bleeding can be detected within individual frames, when the video is paused, and with no need to reference adjacent frames. Temporal artifacts such as flickering, jerkiness and floating can be noticed while the video is being played [101]. Jana *et al.* [24] investigated video artifact evaluation for two-way video conversations in stationary and mobile scenarios using no-reference spatial metrics blocking, blurring, and temporal smoothness. Results showed that blocking and blurring are highly correlated when they are caused by packet loss. However, different coding techniques can perform different in terms of avoiding loss of high frequency components and show less blurring or blocking in different contexts. Silva *et al.* conducted experiments measuring user annoyance caused by different strength combinations of blockiness, blurriness, and packet loss intensity. Disturbances were inserted in video sequences characterized by diverse content and displayed to subjects on a 23 inch monitor [102]. Results showed that subjects were able to identify artifacts only when one source of impairment with high strength was present, while they had difficulties identifying low strength artifacts. A higher level of annoyance correlated with more artifacts being included in the experiment and their respective intensities. Subjects reported that blockiness had the strongest impact on “annoyance”, and in some cases blurriness masked impairments caused by packet loss. The possibility to estimate perceivable quality impairments in terms of blockiness and audio distortion using machine learning, and to predict the occurrence of disturbances was investigated in [103]. The authors studied call scenarios with no impairments and with realistic technical impairments (packet loss and delays). Results showed that impairments could be estimated with a high level of accuracy, thus proving the potential of exploiting machine learning models for automated QoE-driven monitoring and estimation of WebRTC performance.

Layout: Trends have shown increases in smartphone screen sizes, aiming to accommodate higher screen resolutions. High resolution displays impose additional load on the processing unit, particularly on the graphics processor, needed to render high definition images faster. Smartphone screen sizes will most likely not be much bigger in the future, since carrying devices with displays larger than 6” and noticeable weight is not particularly convenient. Thus, an important feature of any service is the possibility to adapt the layout and content to viewing contexts and devices. Even though smartphone displays are relatively small, with limited options for manipulating the design layout, results from studies focusing on desktop video conferencing should be taken into consideration and further extended to mobile devices. Namely, user preferences for layout on single and dual desktop monitors in case of a three-party video conferencing were tested to investigate the relationship with QoE [36]. Authors concluded that directional gaze was not so important to the participants, however, they tried to place the preview windows in a way that preserves directional gaze. Preferred preview window size did not have significant impact but they prefer to keep equal previews. They also preferred the layout where all preview windows can be seen at once without the need to their turn head.

Gunkel *et al.* further studied how packet loss and streams with different video quality impact the layout choice (*fixed layout*: all participants’ previews displayed on the screen, or *layout with changing focus*: the participant currently talking has the biggest preview window) [35]. Authors reported higher loss rates occurred with the high quality streams, which caused more distortions as compared to low quality streams. Nevertheless, overall results showed that participants preferred a fixed layout over a focus-changing layout.

Context influence factors

Quality perception is context-dependent [20], hence the context information (such as mobility, number of participants, time of the day) can provide valuable data that can be further considered in QoE management and in combination with other groups of influence factors. In an audiovisual conversational services, task is recognized as a factor with significant impact on the perceived quality. When considering the *type* of task, participants can be requested to multitask (execute multiple activities simultaneously) or focus on one aspect only [88].

Schmitt *et al.* investigated the impact of video quality on the ability to interact in experiments involving a four-party desktop video conference, where participants were given the task of collaboratively building a Lego model. During the experiment, authors assessed engagement, asking subjects about enjoyment with respect to completing the assigned Lego task. Results showed that subjects with a higher engagement in the task reported a higher QoE [104]. In [105] authors explored the effect of task (with three different level of complexity: simple, moderate, and difficult) and/or duration (30, 60, and 120 seconds) on the overall QoE for WebRTC video call (established over smartphone) using the statistical method analysis of variance. Obtained results confirmed that QoE is significantly determined by the task complexity and duration.

De Moor *et al.* explored the influence of the distorted audio and video due to packet loss, delay and jitter in two-party audiovisual calls established using a WebRTC application [106]. As a conversation incentive, authors used the *Celebrity name guessing* task. Authors concluded that the test task turned out to be more engaging than was intended, impacting the QoE in an unwanted way. Therefore, the task to use during QoE assessment tests has to be chosen carefully so as to avoid situations where subjects are more concentrated on finishing the task then on evaluating the system in terms of different quality features.

Human influence factors

The participant's personality can have a significant influence on perceived quality, especially in determining the conversational structure, in terms of turn-taking behaviour, single-, double- or multi-talk situations. Turn-taking in every day communication usually does not present a problem. However, during a video conversation, due to the delay, participants have been reported as having problems identifying the source of delay-related impairments as being technical (e.g., attributed to network impairments) or interpersonal (behaviour related, e.g., a slow speaker simply makes longer pauses than a fast speaker) [37], [38].

While it is clear that a wide range of system, context, and human IFs affect QoE in multiparty audiovisual telemeetings, questions remain as to the level of the impact of particular factors, especially in a mobile context. For example, the question of whether certain impairments cause strong, noticeable or imperceptible quality degradation commonly depends on the particular scenario, context, as well as the individual involved users.

3.1.2 QoE features/dimensions

Jekosch defined a **quality feature** as *a perceivable, recognized and nameable characteristic of the individual's experience of service which contributes to its quality* [107], where a feature as a characteristic of a perceptual event can be seen as a dimension in perceptual space. In [108], the authors categorized QoE features related to the perceived quality on the following five levels:

Level of direct perception includes quality features related to the perceptual event created immediately and spontaneously during the experience (e.g., audio: timbre, noise, speech intelligibility, coloration, interruptions; video: sharpness, darkness, brightness, flicker, color bleeding, jerkiness, blockiness, blurriness),

Level of action relates to the human perception of his/her own (e.g., audio: sidetone, echo, or double-talk degradation; video: immersion, perception of space and own motions),

Level of interaction involves the exchange of actions and re-actions, human-to-human and human-to-machine interaction (responsiveness, naturalness of interaction, communication efficiency, and conversation effectiveness),

Level of the usage instance of the service relates to physical and social usage situation (e.g., learnability and intuitivity of the service, effectiveness and efficiency in achieving a specific goal).

Level of service relates to the usage of the service beyond a particular instance (e.g., appeal, usefulness, utility, acceptability).

Subsequently, Skowronek proposed three additional group communication levels, first to **establish common ground** (paying attention to what is being said and actively listening, providing verbal and nonverbal feedback), second to describe **group-conversation dynamics** (turn taking, double talk), and third referring to the **cognitive load** (user working memory, with limited processing capabilities, fatigue) [20].

As already mentioned, a large number of factors and features may have an impact on the QoE of multiparty audiovisual telemeeting services. Depending on the available resources, number of participants, and the social context, the importance and particular impact of individual factors will differ. Even though QoE is a multidisciplinary and multidimensional concept, there is a need to narrow down consideration of potential influence factors so as to address key factors to be considered when developing QoE models to be utilized for QoE management purposes.

3.2 Quality assessment

Multiparty telemeetings can differ in various aspects, hence a number of ITU-T and ITU-R Recommendations are used to describe subjective quality evaluation methods, each focusing on individual communication modes, test modes or types of quality, as summarized in Table 3.2.

The main difference between a two-party and a multiparty call lies in the more complex conversational situation resulting in the multiparty case. A set-up with more participants involved in the session allows talking with more than one person simultaneously, creating group

Table 3.2: ITU-T and ITU-R Recommendations addressing quality assessment for conversational services and differing in terms of type of quality, communication mode and test mode.

Type of quality	Communication mode	Test mode	ITU-T/ITU-R Recommendations
Non-interactive	Audio-only	Audio	P.800 [9], P.806 [109], P.880 [13], BS.1116 [15], BS.1285 [16], BS.1534 [110], P.1302 [111], P.1310 [112], P.1311 [113]
Non-interactive	Video-only	Video	P.910 [14], BT.500 [114], BT.710 [115], BT.1788 [116], P.915 [117], P.916 [118]
Non-interactive	Audiovisual	Audio	P.800 [9], P.880 [13], BS.1116 [15], BS.1285 [16], BS.1534 [110], P.913 [119]
Non-interactive	Audiovisual	Video	P.910 [14], BT.500 [114], BT.710 [115], BT.1788 [116], P.913 [119]
Non-interactive	Audiovisual	Audiovisual	P.911 [10], P.1302 [111], P.913 [119]
Conversational/ Interactive	Audio-only	Audio	P.800 [9], P.805 [12], P.1305 [120], P.1310 [112], P.1312 [121]
Conversational/ Interactive	Audiovisual	Audio	P.800 [9], P.805 [12], P.1310 [112], P.1312 [121]
Conversational/ Interactive	Audiovisual	Audiovisual	P.920 [11], P.1305 [120]

dynamics. As a result, the required cognitive load may be different, which may in turn impact quality judgments [20] .

Dimensionality of assessment and rating scales

Multiparty audiovisual telemeetings are commonly assessed using purpose-designed opinion questionnaires, where participants complete and report ratings for audio-visual and overall quality, AV synchronization and/or interactivity degradation. In terms of time frame, quality may be assessed [122]:

- after stimulus/stimuli presentation, or
- continuously during stimulus/stimuli presentation.

Different **rating scales** are used to correlate opinions with numerical values, enabling the calculation of arithmetic mean (in the case of ordinal scales assuming equal intervals between quality levels). Different scales are used depending on the judgment type. If ratings are collected after participants have been exposed to stimuli, then it is common to use a discrete 5-point absolute category rating (ACR) scale for quality marked with: 1 ***"Bad"***, 2 ***"Poor"***, 3 ***"Fair"***, 4 ***"Good"***, 5 ***"Excellent"***, while interactivity degradation is marked with: 1 ***"Very annoying"***, 2 ***"Annoying"***, 3 ***"Slightly annoying"***, 4 ***"Perceptible but not annoying"***, 5 ***"Imperceptible"*** [9]. For instantaneous judgment, a continuous scale with the same labels is suggested. Participants assess quality by moving a slider during the session, where the slider position corresponds to the currently perceived quality level.

It is a common to use Mean Opinion Scores (MOS) [123] to quantify perceived quality, by averaging the ratings of all test participants for the same test scenario [124].

Set-up of a multiparty telemeeting assessment test

The recommended conversation time length depends on the number of test participants [17].

If there are only a few participants, about one and a half minutes should be added per participant to obtain a feasible test length, and one minute per participant if there are around six participants. Determining an appropriate test time length is very important, since experiments can last long due to a large number of test conditions. Over time, due to tiredness, test subjects' attention is reduced.

Type of task

Every experimental design decision impacts the user experience, hence it is important to pay attention to the ***type of task*** as well. Test tasks for audiovisual conversational tests are described in ITU-T P.920, but most tasks are designed for two participants [11]. Tasks to evaluate the effects of speech delay on communication quality include :

- take turns in counting,
- take turns reading random numbers aloud as quickly as possible,
- take turns verifying random numbers aloud as quickly as possible,
- words with missing letters are completed with letters supplied by the other talker,
- take turns verifying city names as quickly as possible,
- determine the shape of a figure described verbally, and
- free conversation.

Standards recommend use of the **free conversation** test task, due to the resemblance to real-life natural conversations, enabling participants to keep their focus on the screen. However, there are cases where, due to limitations in the pool of available participants, it may not be possible to group extroverted acquaintances. In such cases, ITU-T Recomm. P.1301 proposes the use of three types of tasks [17]:

- *Survival task* (where participants as a group need to decide which items will help them to survive),
- *Leavitt task* (where participants have to find, among five others, a common object drawn on their paper),
- *Brainstorming task* (where participants have to jointly generate a maximum number of unusual ideas).

Some of the tasks with predetermined scenarios are not considered suitable for audiovisual quality evaluation since subject attention is divided between the paper and display.

3.2.1 Subjective quality assessment

In the context of quality assessment for multiparty telemeetings, an important aspect to consider is the **number of participants** and **number of participant locations** [17]: two sites with more than one person at at least one site (**multiparty point-to-point**), more than two sites with one person at each site (**multiparty one-per-site**), and more than two sites with more than one person at at least one site (**multiparty multi-point**), as shown in Figure 3.2.

Another important aspect of the telemeeting system that has to be considered is **communication mode** which may refer to **audio-only, video-only, or audiovisual** mode. Sign language communication uses hand gestures, facial expressions and signs, and for such communication video only mode can be used. Communication mode can be rendered differently. Sound can be reproduced using mono channel or spatial technology, while video can be displayed as non-spatial 2D or spatial 3D. Applications based on the WebRTC technology [125], enable data transmission along with media streams, providing additional **text** (e.g., email, chat) and **graphic** (e.g., pictures, slides) based information exchange.

Evaluation mode and type of quality

Telemeetings can further differ in terms of the type of quality dimension. *Non-interactive quality* may be assessed by listening-, viewing-, or listening-and-viewing-only quality of test stimuli, while *conversational/interactive* quality is commonly assessed by participants engaged in an actual conversation [17].

Two categories of interactivity are identified, interactivity as a process or as a product [126]. Interactivity as process *is interaction taking place between human subjects where subsequent messages consist of responses to prior messages or requests in a coherent fashion. Roles of the interactants (humans, machine, media) are in general reciprocal and can be exchanged freely*. Interactivity as a product *occurs when a set of technological features allows users to interact with the system*. In multiparty telemeetings, the ability to interact has a significant impact on the QoE. The interactivity process is comprised of IFs, interaction performance aspects and perceived quality features (Figure 3.3 [126].

Accordingly, a telemeeting is a combination of the type of quality, communication and test mode (based on audio-only, video-only, or audiovisual quality). For example, the communication mode can be audiovisual, but only video quality will be evaluated.

Controlled and non-controlled environments

Evaluation can take place in a laboratory environment under controlled conditions, or in

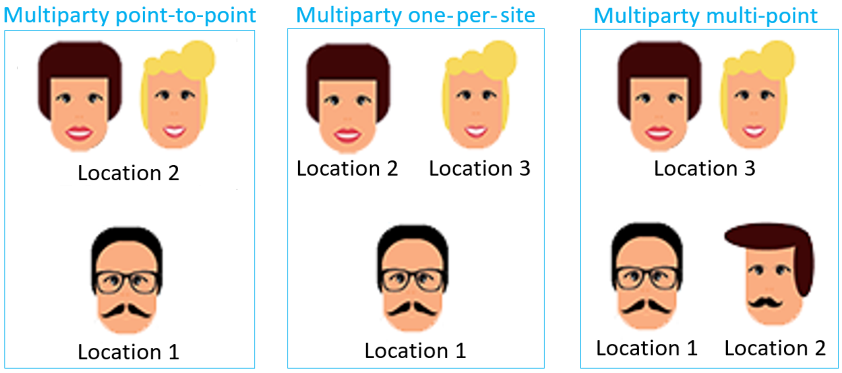


Figure 3.2: Minimum multiparty setup defined by number of participants and number of locations.

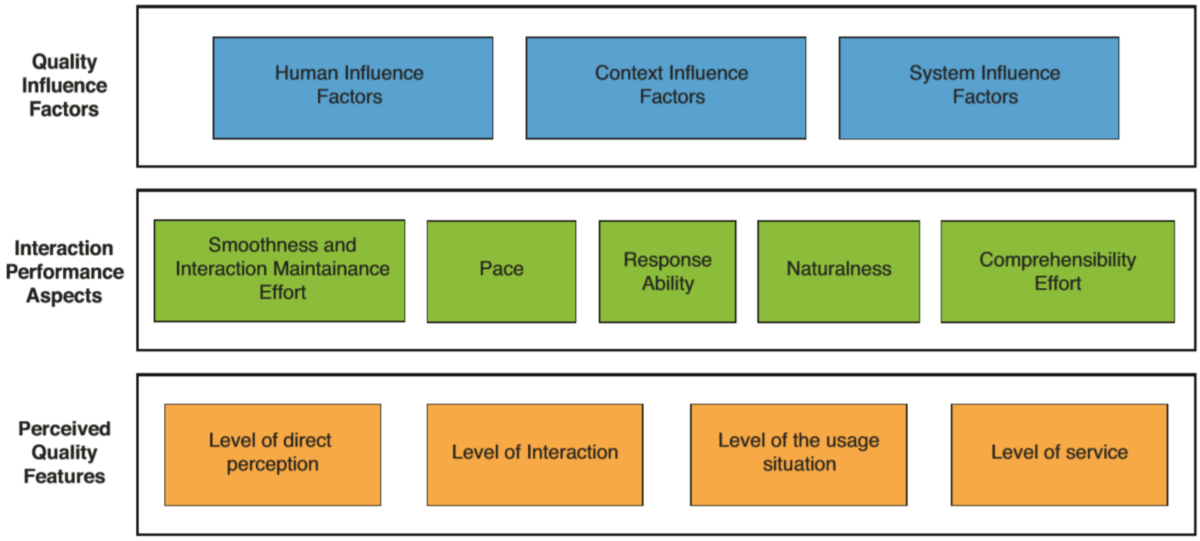

Quality Influence Factors
Human Influence Factors
Context Influence Factors
System Influence Factors
Interaction Performance Aspects
Smoothness and Interaction Maintainance Effort
Pace
Response Ability
Naturalness
Comprehensibility Effort
Perceived Quality Features
Level of direct perception
Level of Interaction
Level of the usage situation
Level of service

Figure 3.3: Interaction performance aspects and quality features [126].

real-life situations and non-controlled environments, in which case the environment should be well described and specified.

Symmetric and asymmetric set-up

Finally, an important consideration is the system set-up. In a multiparty environment, participants may be using heterogeneous devices and access networks. Consequently, they may not only experience different impairments and quality degradation, but may also have different quality expectations. As an example, we can envisage a real-life scenario involving two participants with high-end PCs, large screens, and connected to high speed fixed networks in their offices, communicating with a third participant who is using a mobile phone with a 5.1" display, traveling on a train, and has a poor mobile network connection.

The potentially high diversity within a multiparty service scenario makes it difficult to use general evaluation methods for all types of equipment used in all circumstances. Even though certain impairments are not present at each site, participants can be aware of asymmetric disturbances due to two or more previews of other participants on the screen. In their work reported in [22], the authors found that asymmetric impairments, even with one degradation source, multiply perceptible disturbances, resulting with different perception by each user.

Multimedia quality models

The objective perceptual multimedia quality model described in ITU-T Recomm. J.148 [127] includes three input modules, providing predictions of audio and video quality for a multimedia service, and an indication of the differential delay between the audio and video signals (Figure 3.4). An additional task module serves as input for the multimedia integration function capturing the human perceptual and cognitive processes in the formation of quality judgments. Task presents the degree of interactivity associated with the multimedia service. The goal is to define a set of integration rules that enable the multimedia model to accurately predict perceived

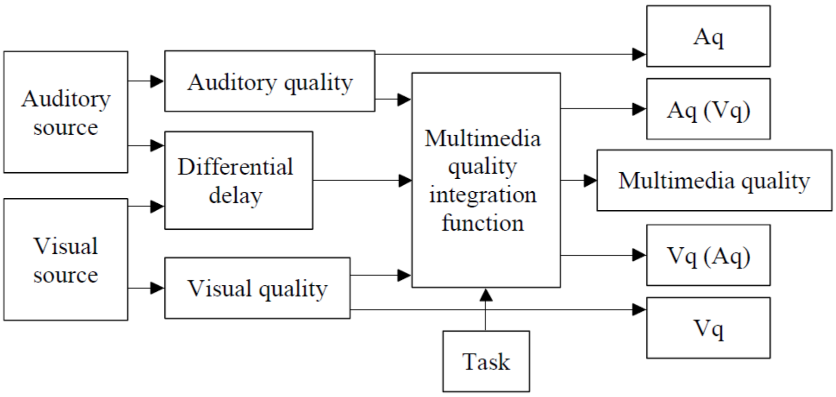


Figure 3.4: Basic components of a multimedia model [127].

quality of multimedia services.

The Recommendation defines multimedia as *the combination of multiple forms of media such as audio, video, text, graphics, fax, and telephony in the communication of information*, whereas *Aq* presents objective measurement of audio quality, and *Vq* objective measurement of video quality. The component *Aq(Vq)* is an objective measurement of audio quality, accounting for the influence of video quality, while *Vq(Aq)* represents an objective measurement of video quality, accounting for the influence of audio quality.

Audiovisual quality integration

In [128] the authors propose a general audiovisual quality model based on the late fusion theory, where audio and video qualities are fused at the late stage of the overall perceived quality judgment process [129]. The proposed model involves two predictors, audio (MOS_A) and video *(MOS_V)* quality dimensions, and corresponding *A, B, C*, and *D* scalar coefficients (eq. 3.1):

$$MOS_{AV} = A \cdot MOS_A + B \cdot MOS_V + C \cdot MOS_A \cdot MOS_V + D \tag{3.1}$$

The authors proposed two models for video telephony based on interactive experiments, the first based on short conversation tests (SCT) (eq. 3.2), and the second based on audiovisual short conversation tests (AVSCT) (eq. 3.3) simulating an ‘average’ video call with a leveled use of the audio and video channels, consisting of a semi-structured dialog where participants answer each other’s questions.

$$MOS_{AV-SCT} = 0.548 \cdot MOS_A + 0.357 \cdot MOS_V + 0.0013 \cdot MOS_A \cdot MOS_V + 0.127 \tag{3.2}$$

$$MOS_{AV-AVSCT} = 0.03 \cdot MOS_A + 0.079 \cdot MOS_V + 0.123 \cdot MOS_A \cdot MOS_V + 1.374 \tag{3.3}$$

Audio-video synchronization quality model

In [130] Saidi *et al.* proposed an audio-video synchronization quality prediction model (eq. 3.4): for video communication based on the perceived audio quality, perceived video quality, and the desynchronization predictor $DMOS_{synch}$:

$$MOS_{AV} = 1.57 + 0.16 \cdot MOS_A * MOS_V - 0.15 \cdot DMOS_{synch} \tag{3.4}$$

where $DMOS_{synch}$ is calculated as (eq. 3.5):

$$DMOS_{synch} = 5 - MOS_{\text{synch}} \tag{3.5}$$

MOS_{synch} rating evaluates synchronization determined by the 5 point scale: 1 *"Very annoying"* to 5 *"Imperceptible"*.

Opinion model for quality assessment in videotelephony

ITU-T Recommendation G.1070 defines a procedure that estimates media quality (taking interactivity into consideration) for videotelephony applications. The model is based on three functions: video quality estimation, speech quality estimation, and multimedia quality integration (Figure 3.5). Additionally, the multimedia quality integration function takes into consideration an end-to-end delay. The model outputs multimedia quality (MM_q), video quality influenced by speech quality (*Vq(Sq)*), and speech quality influenced by video quality (*Sq(Vq)*).

The MM_q (eq. 3.6) is calculated using audiovisual quality MM_{SV} and audiovisual delay impairment factor MM_T, which represents the degree of the audiovisual quality degradation due to audiovisual delay and synchronization.

$$MM_q = m_1 \cdot MM_{SV} + m_2 \cdot MM_T + m_3 MM_{SV} \cdot MM_T + m_4 \tag{3.6}$$

Coefficients m_1 to m_4 are dependent on video display size and conversational task.MM_q ranges from 1 to 5.

Video quality is modeled as the product between I_{coding} (eq. 3.8) basic video quality affected by the coding impairments (under given video bitrate and video frame rate), and the packet loss degradation. D_{PplV} (eq. 3.12) defines the degree of video quality robustness against packet loss and P_{plV} *[%]* video packet-loss rate.

$$V_q = 1 + I_{coding} \cdot exp(-\frac{P_{plv}}{D_{PplV}}) \tag{3.7}$$

The basic video quality $\mathrm{I}_{\mathrm{coding}}$ (eq. 3.8) impacted by coding impairments is defined as:

$$I_{coding} = I_{Ofr} \cdot exp(-\frac{(ln(Fr_V) - ln(O_f r))^2}{2D_{FrV}^2}) \tag{3.8}$$

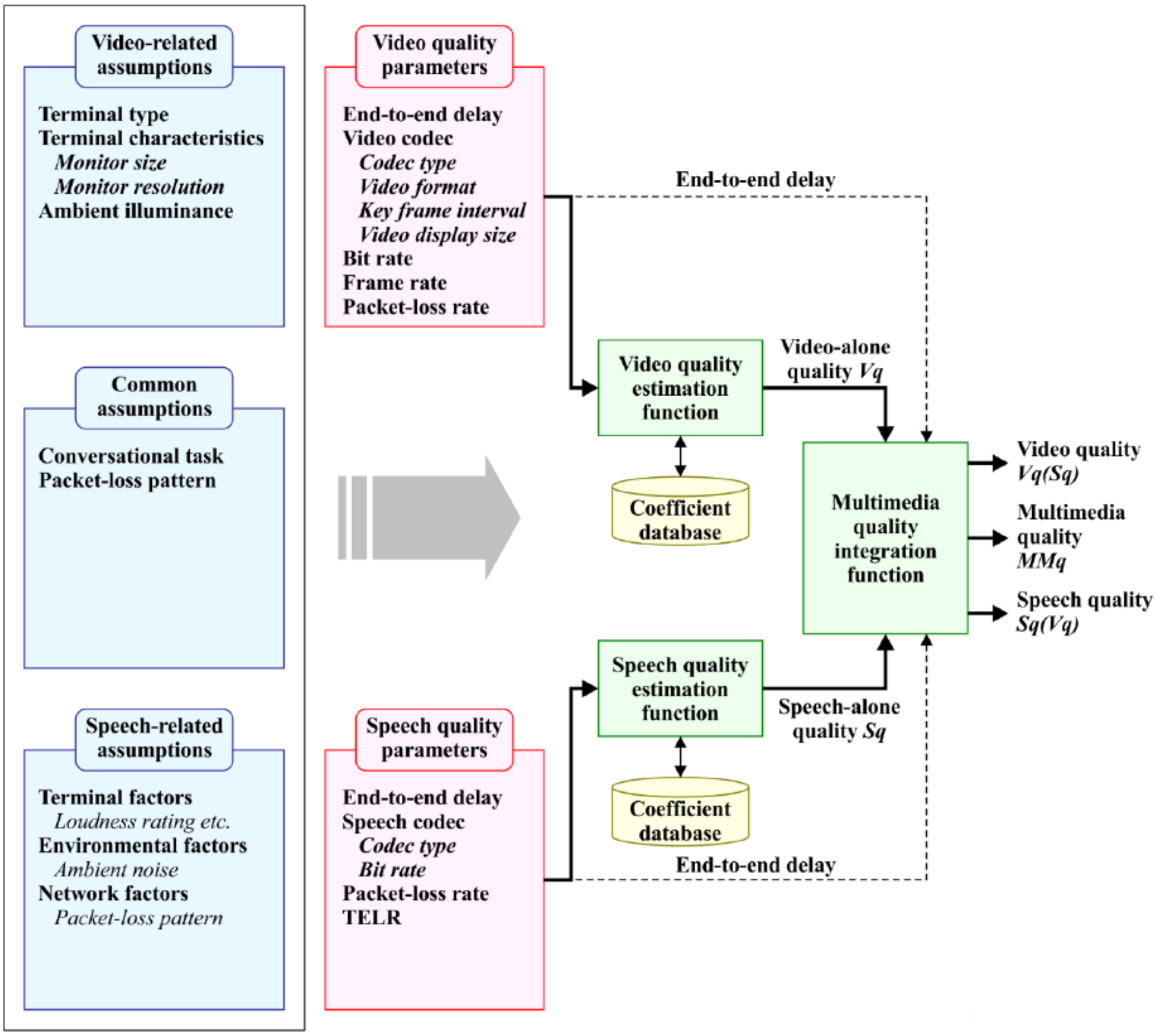


Figure 3.5: Diagram of a multimedia communication quality assessment model [131].

where D_{frV} (eq. 3.11) denotes degree of video quality robustness due to frame rate reduction (Fr_V), and O_{fr} (eq. 3.9) optimal frame rate that maximizes video quality at each bit rate (Br_V). O_{fr} is defined as:

$$O_{fr} = v_1 + v_2 \cdot Br_V \tag{3.9}$$

with $0 <= O_{fr} <= 30$; v_1 and v_2: const. Where $F_{rV} = O_{fr}$, $I_{coding} = I_{Ofr}$, I_{Ofr} (eq. 3.10) represents objective measurement of maximum video quality at each bit rate (Br_V):

$$I_{Ofr} = v_3 - \frac{v_3}{1 + (\frac{Br_V}{v_4})^v_5} \tag{3.10}$$

with $0 <= I_{Ofr} <= 4;$ v_3, v_4, and v_5: const.

$$D_{frV} = v_6 + v_7 \cdot Br_V \tag{3.11}$$

with $0 < D_{FrV};$ v_6 and v_7: const.

The packet loss robustness factor D_{PplV} is defined as:

$$D_{PplV} = v_{10} + v_{11} \cdot exp(-\frac{Fr_V}{V_8}) + v_{12} \cdot exp(-\frac{Br_V}{V_9}) \tag{3.12}$$

with $0 < D_{PplV}$.

Coefficients v_1 to v_{12} are dependent on the codec type, video format, key frame interval and video display size. An inherent limitation of a model is that subjective opinions are necessary before model can be used. Namely, coefficients are dependent on the specific scenario and have to be determined prior to the model usage.

3.3 Overview of studies addressing audiovisual telemeetings QoE

Table 3.3 summarizes existing literature in domain of QoE assessment for video conferencing services, addressing study specific parameters such as environment where the assessment test were conducted, used application, number of participants included in the study, selected system, context and human influence factors, type of end user device, type of quality and rating scale. Lastly, the table contains the key findings of each study.

Over the past few years, substantial progress in enhancing QoE have been made, addressing complex and multi-faceted challenges of services such as audiovisual telemeetings, identifying relationships between end-user QoE and various network, service, and context factors. Commonly considered system influence factors in terms of network quality were delay, jitter and packet loss, while in terms of video quality were video resolution, bitrate, and frame rate. Rao *et al.* explored bandwidth, jitter, and packet loss in experiments comparing recorded video files with the reference videos [132]. Obtained results confirmed that bandwidth is directly proportional to the perceived quality, while frequent bandwidth fluctuations quickly degrade perceived quality. While a single packet loss or short jitter impairment did not impact QoE severely, combined impairments were found to significantly impacted QoE. Balihodžić *et al.* conducted experiments over laptop and smartphones examining the impact of CPU, delay, and jitter on the QoE of WebRTC video calls [133]. Results showed that CPU had the weakest impact on QoE, while delay and jitter had a negative medium and strong correlation with QoE, respectively. In a similar study, authors identified the CPU impact of the end user device on the delay of call establishment and video quality [134]. Even though video applications can use multiple cores available in low-end phones, video calls are linearly affected by slower CPU speeds mainly because of the packet processing overhead and partly due to the TCP processing delays.

It is well known that video quality significantly impacts perceived quality, especially in a multiparty setup where different conditions per participant can include different previewed quality levels on the receiver side. In [135], authors have shown that in a multiparty setup, participants differently rate overall screen quality (encompassing all video streams portrayed simultaneously on the screen) as compared to the mean quality of each individual video stream. Results also showed that overall quality can be significantly enhanced if only one stream of a low quality is replaced with the high quality one. When it comes to contextual QoE influence factors, different types of tasks were the most researched. In [136] authors investigated the individual and combined impact on the QoE and its dimensions (overall satisfaction, efficiency, ease of use, acceptance) of task type and application type. Experiments involved three-party video conference calls via smartphones and using WebRTC. Authors concluded that the task type has no significant impact on QoE and ease of use. This is in contrast to other QoE dimensions (satisfaction, efficiency, and acceptance) where type of task showed significant impact. Finally, in study [96] that involved 32 different system and context influence factors along with five human IFs, such as difficulties in using modern technologies, emotional state, hand injuries, visual difficulties and speech difficulties, authors investigated the impact of the selected factors on QoE. Although the impact of HIFs is unquestionable and according to the previous research HIFs as a predictors could be accounted for 24.5% of overall experience, results showed that in a video conference setup, the examined HIFs had the lowest impact [137].

Author (Year)	**Application/ environment**	**No. of participants (if applicable)**	**System IFs**	**Context IFs**	**Human IFs**	**End user device**	**Type of quality**	**Rating scale**	**Multiparty setup**	**Key findings**
Karadža *et al.* (2020) [136]	WebRTC/ natural	18	-	Task, used app	-	Smartphone	Interactive	5-point Likert	Yes	Proposed Multi-Layer Perception Artificial Neural Network prediction model based on overall satisfaction, efficiency, ease of use, and acceptance.
Balihodžić *et al.* (2020) [133]	WebRTC/ natural	20	Delay, jitter, CPU usage	-	-	Laptop/ smartphone	Interactive	5-point ACR	No	Jitter is strongly negatively correlated with the QoE, while delay had a medium negative, and CPU weak and positive correlation with QoE.
Rao *et al.* (2019) [132]	Video clips WebRTC/ laboratory	24	Delay, jitter, packet loss, download rate, upload rate	-	-	-	Non-interactive	DMOS	No	Quantified relationship between the main QoS parameters (bandwidth, jitter and packet loss) and perceived video quality.
Baraković Husić *et al.* (2019) [105]	WebRTC/ natural	30		Task complexity and duration (guessing shapes)	-	Smartphone	Interactive	5-point ACR	No	Proposed QoE prediction linear model, with complexity and duration as a predictors.
Baraković Husić *et al.* (2018) [96]	Online questionnaire/ natural	140	Audio quality, image quality, loss of video frames, price, ease of use, procedure of accessing web environment	Time of the day	Difficulties in using modern technologies, emotional state, hand injuries, speech and visual difficulties	Smartphone	Non-interactive	5-point ACR	No	Identification of most influential factors for a video calls.

Author (Year)	**Application/ environment**	**No. of participants (if applicable)**	**System IFs**	**Context IFs**	**Human IFs**	**End user device**	**Type of quality**	**Rating scale**	**Multiparty setup**	**Key findings**
Schmitt *et al.* (2018) [135]	Video clips/ crowdsourcing study	5000	Bitrate	-	-	-	Non-interactive	5-point ACR	Yes	Co-present mixed qualities in a multiparty setup impacts the perceived quality, where users are more critical when asked for individual streams than for an overall rating.
Garcia *et al.* (2018) [97]	Literature review	N/A	-	-	-	-	-	-	No	Proposed KPIs for WebRTC QoE estimation: call establishment time, end-to-end delay, audio quality, video quality, and audiovisual quality.
Dasari *et al.* (2018) [134]	Video clips (Skype)/ laboratory	N/A	Delay, frame rate	-		Mobile device, laptop	Non-interactive	-	No	Identification of the CPU impact on the delay of call establishment procedure and video quality.
He *et al.* (2018) [138]	Video clips (Skype)/ laboratory	24	Bitrate, frame rate	-		Mobile device, cloud based device	Non-interactive	-	No	Proposed scalable QoE model based on passive network measurements (packet size, inter-packet arrival time, TCP bytes in flight, number of received packets, and packet loss.
Schmitt *et al.* (2017) [104]	QoE-TB/ natural	28	Bitrate, packet loss	Building blocks task	-	Desktop	Interactive	5-point ACR	Yes	QoE of more engaged participants is higher than that of the less engaged participants. Low bitrates affect significantly the interaction, impacting the movement patterns of users as well as speech patterns.

Author (Year)	Application/ environment	No. of participants (if applicable)	System IFs	Context IFs	Human IFs	End user device	Type of quality	Rating scale	Multiparty setup	Key findings
De Moor *et al.* (2017) [30]	WebRTC/ natural	22	Video packet loss, AV delay, AV jitter, CPU usage	Celebrity name guessing task	-	Smartphone, laptop	Interactive	9-point Self-Assessment Manikin, and 5-point ACR	No	Varying quality impacts overall quality and annoyance with a lower tolerance for audio distortions. Distortion of audio and video leads to, lowest quality and highest annoyance ratings.
Jana *et al.* (2016) [24]	Video clips (Vtok, Skype)/ natural	24	Delay, packet loss, bandwidth	Task complexity (guessing shapes)	-	Smartphone preview on the computer	Non-interactive	5-point ACR	No	Proposed QoE prediction model based on end-user movement, network delay, loss rate and bandwidth.
Gunkel *et al.* (2015) [35]	Vconect/ natural	20	Video resolution, packet loss, layout	Survival task	-	Desktop	Interactive	9-point likert-like	Yes	Packet loss and the resulting distortions have a greater impact on the QoE than reduced video quality due to lowered resolution.
Skowronek *et al.* (2013) [22]	Conference bridge (Asterisk)/ laboratory	51	Packet loss, echo	Background noise, loudness	-	VoIP-telephones (SNOM870)	Interactive	5-point ACR	Yes	Proposed a systematic method for multiparty setup with asymmetric impairments, describing how individual technical degradations impact perception of other participants.
Vakili *et al.* (2013) [32]	Video clips/ laboratory	25	Frame rate, quantization parameter (compression level)	-	-	15" laptop	Non-interactive	11-level ACR	No	Proposed QoE control mechanism based on bandwidth limitation, frame rate and compression level.

Table 3.3: An overview of conducted QoE studies involving audiovisual telemeetings.

3.4 Chapter summary

This chapter explains the concept of QoE, and discusses challenges related to the subjective quality assessment of multiparty telemeetings. Narrowing down to the thesis focus, we give a thorough state-of-the-art analysis of conducted research on QoE for audiovisual telemeetings. Numerous influence factors investigated in conducted studies involving audiovisual telemeetings have shown to have an impact on the final quality judgment. Results showed that providing high levels of QoE for audiovisual telemeetings calls for the need to address challenges at both network and application layers. Given that our focus in the scope of this thesis is on multiparty video calls established in a leisure context, we aimed to obtain insights into what end users consider to be the most influential factors in such a context. The following chapter thus reports on the results of an extensive online survey conducted to obtain feedback with respect to IFs and user expectations.

Chapter 4

End user survey addressing multiparty audiovisual telemeeting QoE IFs

Given the wide range of human and context factors that may impact end user expectations and quality ratings, we conducted a web-based questionnaire survey with the goal being to investigate users' opinions and expectations related to audiovisual telemeetings on mobile devices and focusing on the leisure/private context. The aim of this questionnaire was to investigate the factors that subjects identify as most influential in contributing to their overall experience and quality perception. The questionnaire was written in the Croatian language, and its English translation is provided in Appendix A.

The questionnaire covers ratings of the impacts and importance of considered factors referring to the application, resources, and context. Selected factors belong to the quality features that are possible for a wider audience to evaluate from a perceptual perspective. Questions were divided into the following four groups, with responses analyzed in Sections 4.2 - 4.4:

- **general information** - referring to the subject's demographic data,
- **media quality** - referring to the quality of the sound (audio) and the image (video) in terms of perceivable impairments (e.g., delay, blurriness),
- **functional completeness** - referring to the additional functionalities supported by telemeeting/conferencing services, beyond only audiovisual calls, and
- **service quality and usability** - referring to the ease of use of the application, and the extent to which users feel they are able to conduct audiovisual calls.

To gather user feedback on the perceived quality of audiovisual telemeetings, two aspects of service delivery were considered: call initiation, and service operation once the audiovisual call is established. Both aspects are comprised of multiple dimensions that contribute to the overall QoE: effort required by the user, responsiveness of the service, fidelity of information, security, and availability.

The survey was prepared using the Google Docs service (https://docs.google.com) and distributed over email to acquaintances, colleagues and students. 272 participants successfully completed the questionnaire. Responses were collected over a period of thirteen days (from February 13, until February 26, 2020), and included 41 questions. We note that the survey was conducted just prior to the global outbreak of the COVID-19 pandemic, and reflects the views and opinions of users at that time. Given the drastic increase in video communication services resulting from lockdown measures [2], it would be interesting to repeat the study in future work so as to assess whether or not there are any significant differences in user opinions. We offered participants only closed-ended questions that provided a fixed set of options to choose from. Closed-ended response choices were comprised of yes/no options, multiple choice options, and rating scales. Results are described and analyzed in the following sections.

To identify the factors considered by users to be most influential in impacting the user experience and service quality, participants were asked specific questions about their use of audiovisual telemeetings, and specific opinion questions about the IFs. Statistical analysis of collected data was performed using IBM's SPSS [1], whereby we report on percentages and descriptive statistics such as measures of central tendency and variability.

4.1 Participant demographics and previous experience

Common demographic information about age, gender, education, and country of origin were collected to better understand certain background characteristics of users. Users were divided into different age groups. The majority of users (49.6%) fit into the category 36-45 years. The second largest group fit into the 26-35 years category with a share of 25%. Groups 18-25 years and 46-55 years were distributed with the same share of 11%. Only 3.4% of the users were older adults (>55 years). Of the 272 participants, 51.1% reported their gender as female and 48.9% as male. The response rate for educational level high school degree was 19.1%, University degree (bachelor or masters) was 71%, and 9.9% for higher University degree (PhD) in total. A total of 87 males and 106 females reported having a University degree, while 15 males and 12 females reported PhD degrees. The response distribution is shown on the Figure 4.1. The majority of participants, 89%, reported their country of origin as Croatia, while the rest of the participants were from Bosnia and Herzegovina (4.41%), Serbia and Montenegro (in total of 6.6%). Corrected (by contact lenses or glasses) visual impairment was reported by 46.3% participants.

Following the collection of demographic data, the aim of the following set of questions (6-8) was to identify users' habits associated with audiovisual calls. The question addressing device usage was a multiple choice question type that allows participants to select one or multiple answers from a defined list. Out of 272 participants, 94.9% reported a smartphone as the device

[1] https://www.ibm.com/analytics/spss-statistics-software

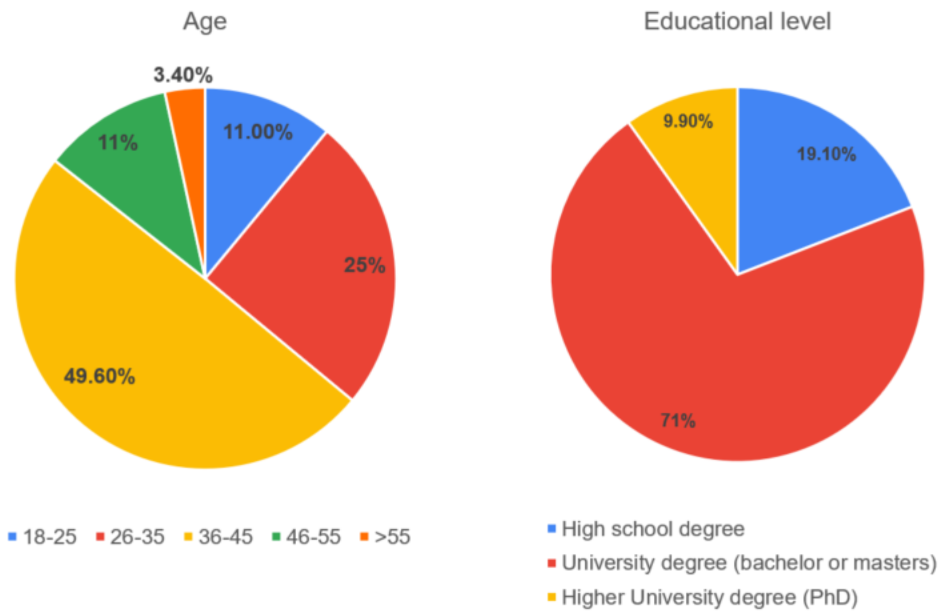


Figure 4.1: Percentage of age and educational level of participants.

used to make video calls, 12.1% reported using a tablet, 52.2% computer/laptop, while 1.5% responded they used some other device (Figure 4.2).

Participants were allowed to provide multiple answers for previously used applications as well. Whatsapp and Skype were the two most commonly used applications in the video call context, with a share of 89.3% and 85.7%, respectively. This was followed by Viber (70.6%) and Google Hangouts (22.8%). Appear.in (renamed to Whereby in 2019) was used by 3.7%, while 26.5% of participants used some other video call app (Figure 4.3).

Of the 272 participants, 16.2% reported participation in video calls in the last 30 days on a daily basis, 14.7% frequently, 16.9% occasionally, 34.6% rarely, and 17.6% never (Figure 4.4). Addressing the multiparty context, 49.3% of users responded positive to having previously participated in a video call with more than two users.

4.2 Opinions related to media quality

To determine key influence factors (according to users, the most important factors or the ones impacting quality and the experience with the greatest extent), we have taken into consideration factors conform two conditions: coefficient of variation is under 27% and mean is higher than 3.6.

From the 10th question until the end of the questionnaire, questions were explicitly related to calls established via a mobile smartphone in a leisure context. Questions regarding the importance of media quality factors and impact on the overall perceived quality of audiovisual call were comprised of closed-ended questions including a predefined list of five answer options.

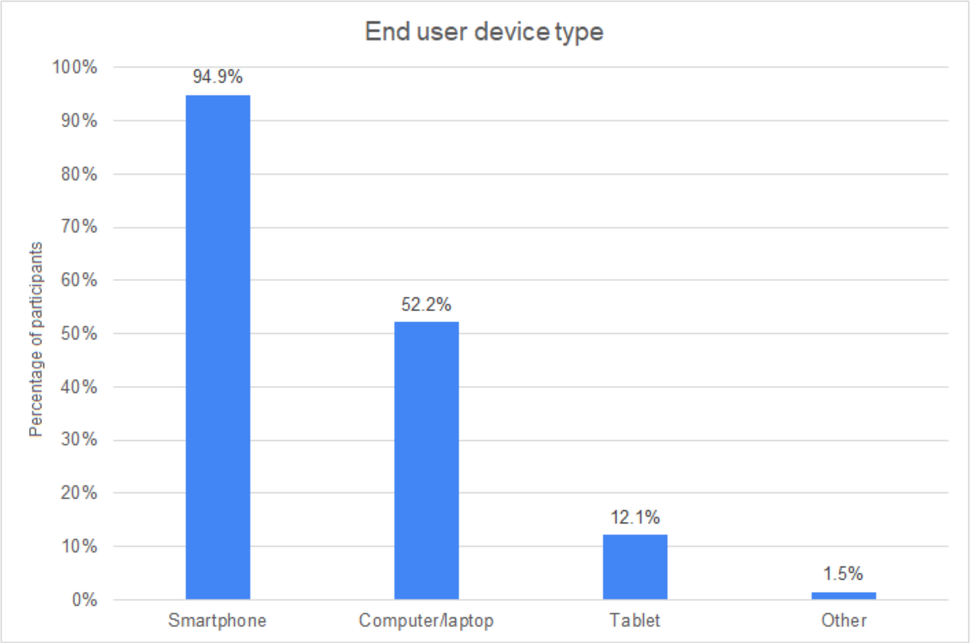


Figure 4.2: Percentage of participants that reported using a given device when making a video call.

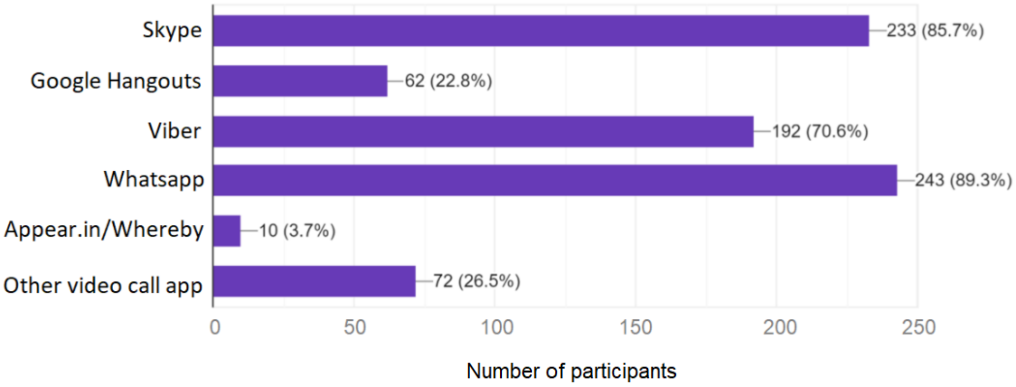


Figure 4.3: Percentage of applications used when making a video call.

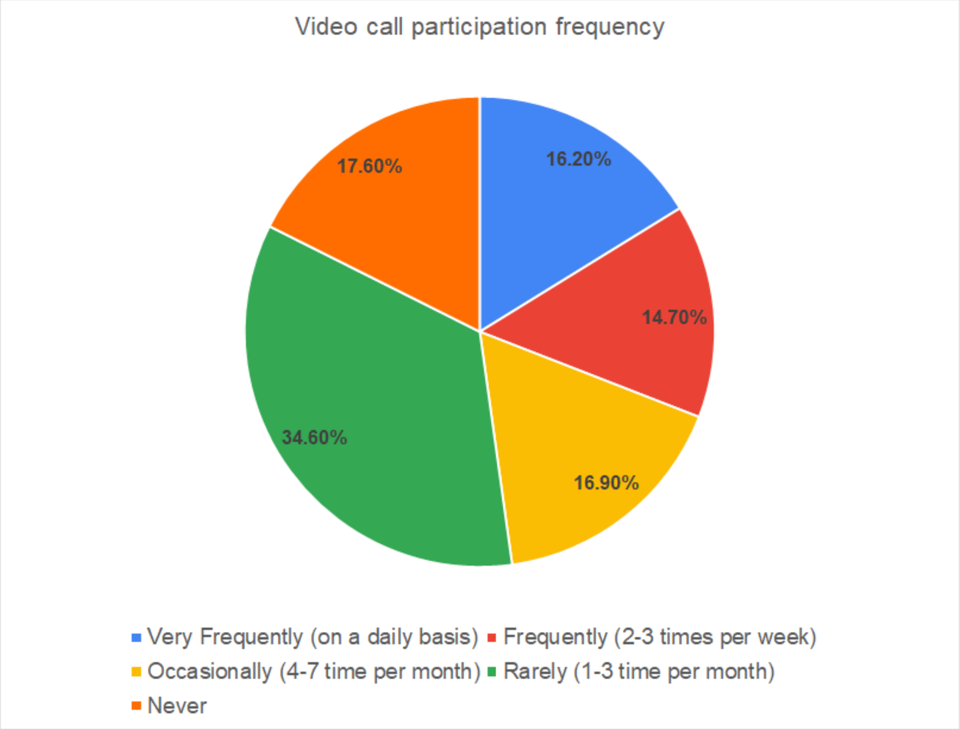


Figure 4.4: Frequency of users that reported having participated in a video call during the last 30 days.

The impact of each factor was rated on 5-point scale. One set of questions used the following rating scale: 5 - “Very Important”, 4 - “Important”, 3 - “Moderately Important”, 2 - “Slightly Important”, 1 - “Not Important”. The other set of questions used the following scale: 5 - “To a great extent”, 4 - “To a moderate extent”, 3 - “To some extent”, 2 - “To a small extent”, 1 - “Not at All”. Descriptive statistics are used to calculate, describe, and summarize collected ratings (Table 4.1).

In accordance with defined conditions for identifying influence factors, from the *Media quality* segment, the following factors are chosen by users in descending order from those considered to be most to the least influential:

- speech intelligibility,
- audio-video synchronization,
- longer video freezes (i.e., longer than 15 seconds),
- perceptible audio delay,
- image blurriness,
- image sharpness,
- and perceptible video delay
- uninterrupted interaction,
- smooth movement,
- voice naturalness.

4.3 Opinions related to service functionality

Questions regarding additional functionalities available during the audiovisual call and corresponding importance were comprised of closed-ended questions including a predefined list of five answer options. The importance of each function has been rated with 5 - “Very Important”, 4 - “Important”, 3 - “Moderately Important”, 2 - “Slightly Important”, 1 - “Not Important”. The descriptive statistics for the collected ratings are given in Table 4.2.

Identified key factors in the questionnaire *Functional completeness* segment in descending order from most influential to least influential are as follows:

- video call recording,
- audio mute,
- adaptive layout (e.g., movable participant’s preview window, display zooming),
- video pause.

4.4 Opinions related to service quality and usability

Questions related to the mobile context, encompassing usability, portability (in terms of efficiency with which the audiovisual call application can be transferred from one operational or usage environment to another), and resource consumption (battery consumption and CPU utilization) were designed as closed-ended questions including a predefined list of five answer options. The importance of each function was once again rated from 5 - “Very Important” to 1 - “Not Important”. The descriptive statistics for the collected ratings are given in Table 4.3.

Factors with the highest impact identified in the scope of the *Usability* segment are the following:

- duration of connection time when establishing a call,
- smooth simultaneous use of other applications (due to the CPU utilization during the call),
- low battery consumption during the call,
- service price,
- security in terms of privacy (i.e., information transmitted during the call is encrypted),
- ease of use of the application,
- user interface aesthetics,
- installation complexity.

Table 4.1: Descriptive statistics for reported ratings (user opinions on media quality factors impacting overall quality) on a scale of 1-5 (1 - “Not Important”, 2 - “Slightly Important”, 3 - “Moderately Important”, 4 - “Important”, 5 - “Very Important” and 1 - “Not at All”, 2 - “To a small extent”, 3 - “To some extent”, 4 - “To a moderate extent”, 5 - “To a great extent”) and frequency of ratings for each influence factor.

Influence factors	Mean	Median	Variance	Standard deviation	Coefficient of variation (%)	Freq. of 1	Freq. of 2	Freq. of 3	Freq. of 4	Freq. of 5	% of 1	% of 2	% of 3	% of 4	% of 5
Speech intelligibility	4.695	5	0.434	0.659	14.036	2	3	9	48	210	0.735	1.103	3.309	17.647	77.206
Voice naturalness	3.702	4	0.797	0.89	24.052	4	12	61	122	73	1.471	4.412	22.426	44.853	26.838
Uninterrupted interaction	3.912	5	0.411	0.893	22.817	2	1	9	72	188	0.735	0.368	3.309	26.471	69.118
Audio-video synchronization	4.629	4	0.741	0.641	13.858	3	8	34	104	123	1.103	2.941	12.5	38.235	45.221
Image sharpness	4.235	4	0.753	0.861	20.331	5	7	80	120	60	1.838	2.574	29.412	44.118	22.059
Smooth movement	3.82	4	0.793	0.868	22.723	3	15	97	102	55	1.103	5.515	35.662	37.5	20.221
Color accuracy	3.276	3	0.99	0.995	30.376	11	42	112	75	32	4.044	15.441	41.176	27.574	11.765
Perceptible audio delay	4.57	5	0.497	0.705	15.426	2	2	16	71	181	0.735	0.735	5.882	26.103	66.544
Perceptible video delay	3.93	4	0.754	0.884	22.487	5	5	30	102	130	1.838	1.838	11.029	37.5	47.794
Image blurriness	4.276	4	0.781	0.868	20.308	4	10	62	121	75	1.471	3.676	22.794	44.485	27.574
Short video freezes	3.489	4	0.745	1.127	32.291	3	11	24	109	125	1.103	4.044	8.824	40.074	45.956
Longer video freezes	4.581	5	0.687	0.829	18.096	5	4	18	46	199	1.838	1.471	6.618	16.912	73.162

Table 4.2: Descriptive statistics for reported ratings (user opinions on functional completeness factors impacting overall quality on a scale of 1-5 (1 - “Not Important”, 2 - “Slightly Important”, 3 - “Moderately Important”, 4 - “Important”, 5 - “Very Important”) and frequency of ratings for each influence factor.

Influence factors	**Mean**	**Median**	**Variance**	**Standard deviation**	**Coefficient of variation (%)**	**Freq. of 1**	**Freq. of 2**	**Freq. of 3**	**Freq. of 4**	**Freq. of 5**	% of 1	**% of 2**	**% of 3**	**% of 4**	**% of 5**
File transfer	3.452	4	1.269	1.055	30.547	15	36	79	85	57	5.515	13.235	29.044	31.25	20.956
Texting	3.294	4	1.112	1.074	32.597	13	34	86	95	44	4.779	12.5	31.618	34.926	16.176
Active speaker identification	3.268	4	0.777	1.161	35.528	3	12	38	118	101	1.103	4.412	13.971	43.382	37.132
Applying make-up / filters/overlay items	3.64	2	1.383	1.21	33.241	127	58	55	19	13	46.691	21.324	20.221	6.985	4.779
Adaptive layout	4	3	1.153	0.901	22.525	19	38	92	90	33	6.985	13.971	33.824	33.088	12.132
Video pause	3.783	3	1.173	0.93	24.573	24	42	112	64	30	8.824	15.441	41.176	23.529	11.029
Audio mute	4.092	4	1.464	0.946	23.127	20	25	70	75	82	7.353	9.191	25.735	27.574	30.147
Call recording	4.257	3	1.348	0.801	18.821	24	41	88	76	43	8.824	15.074	32.353	27.941	15.809

Table 4.3: Descriptive statistics for reported ratings (user opinions on usability factors impacting overall quality on a scale of 1-5 (1 - “Not Important”, 2 - “Slightly Important”, 3 - “Moderately Important”, 4 - “Important”, 5 - “Very Important”) and frequency of ratings for each influence factor.

Influence factors	**Mean**	**Median**	**Variance**	**Standard deviation**	**Coefficient of variation (%)**	**Freq. of 1**	**Freq. of 2**	**Freq. of 3**	**Freq. of 4**	**Freq. of 5**	% of 1	**% of 2**	**% of 3**	**% of 4**	**% of 5**
Interoperability	3.423	4	0.812	1.114	32.55	6	8	50	124	84	2.206	2.941	18.382	45.588	30.882
Connection time	4.449	4	0.864	0.717	16.115	6	17	66	124	59	2.206	6.25	24.265	45.588	21.691
Ease of use	4.033	4	0.642	0.935	23.173	2	8	25	120	117	0.735	2.941	9.191	44.118	43.015
Installation complexity	3.787	4	0.896	1.019	26.921	4	14	44	101	109	1.471	5.147	16.176	37.132	40.074
User interface aesthetics	3.912	4	1.241	0.949	24.251	16	39	79	90	48	5.882	14.338	29.044	33.088	17.647
Reliability of the service	3.125	5	0.514	1.083	34.65	2	3	15	103	149	0.735	1.103	5.515	37.868	54.779
Security in terms of privacy	4.11	5	0.792	0.882	21.452	4	10	21	78	159	1.471	3.676	7.721	28.676	58.456
Low battery consumption	4.39	4	0.873	0.89	20.276	2	17	51	102	100	0.735	6.25	18.75	37.5	36.765
Simultaneous use of other applications	4.415	4	1.039	0.837	18.966	6	27	59	107	73	2.206	9.926	21.691	39.338	26.838
Noise free environment	2.018	4	0.9	1.176	58.274	7	10	63	112	80	2.574	3.676	23.162	41.176	29.412
Price	4.257	5	0.701	0.863	20.278	4	3	29	76	160	1.471	1.103	10.662	27.941	58.824

Calculated frequencies per user rating obtained from the survey results show that factors rated with user rating 4 and 5 together in more than 72% are considered as factors with the highest impact. The following twelve most influential factors in descending order are:

- speech intelligibility,
- audio-video synchronization,
- longer freezes,
- perceptible audio delay,
- low battery consumption,
- image blurriness,
- price,
- security in terms of privacy,
- ease of use,
- perceptible video delay,
- uninterrupted interaction,
- installation complexity.

The last question of the questionnaire aimed to obtain insights into the level of user expectations with respect to quality in terms of free and paid service. More than half of the participants (51.5%) responded they have higher expectations if the service is paid, while 48.5% of participants indicated they have the same level of expectations if the service is free or paid. Participants belonging to the group with the same level of expectations were 54.54% females and 45.46%, while the distribution share of females in the higher level of expectations for a paid service group is 47.86% and 52.14% of males.

With respect to percentage distribution by age group, the most significant difference is within the 25-35 age group where 30% of participants have higher expectations for paid services as compared to 19.69% of participants that have the same level of expectations for both paid and free services (Table 4.4).

The participants were divided into two groups according to their expectations with respect to using a paid service or free service. Group A included participants that stated that they have higher expectations in case of paid services, while group B included participants that stated that they have the same level of expectations in cases of paid and free services. With respect to percentage distribution by user ratings, it was found that 33.92% of all reported ratings by group A corresponded to level 5, while 33.71% corresponded to rating level 4 (Table 4.5). A similar distribution was found in reported ratings by group B where 37.15% of all given ratings was 5 and 33.92% was 4. We can observe that user ratings between those two groups did not differ significantly per age group or per number of given user ratings (observing in total for all impact factors).

Table 4.4: Age percentage of participants with higher expectations for paid services, and participants that reported having the same expectations for paid and free services.

User percentage of participants having		
Age	**higher expectations in case of paid service**	**same expectations in case of paid or free service**
18-25	10.71%	11.36%
26-35	30%	19.69%
36-45	46.43%	53.03%
46-55	9.29%	12.88%
more than 55	3.57%	3.04%

Table 4.5: User ratings combined for all IFs between users with higher expectations for paid services and same expectations for paid and free services.

Percentage of user ratings combined for all IFs within group:		
User rating	**Group A: participants with higher expectations in case of paid service**	**Group B: participants with same expectations in case of paid or free service**
5	33.92%	37.15%
4	33.71%	33.92%
3	20.67%	18.99%
2	7.37%	5.91%
1	4.33%	4.03%

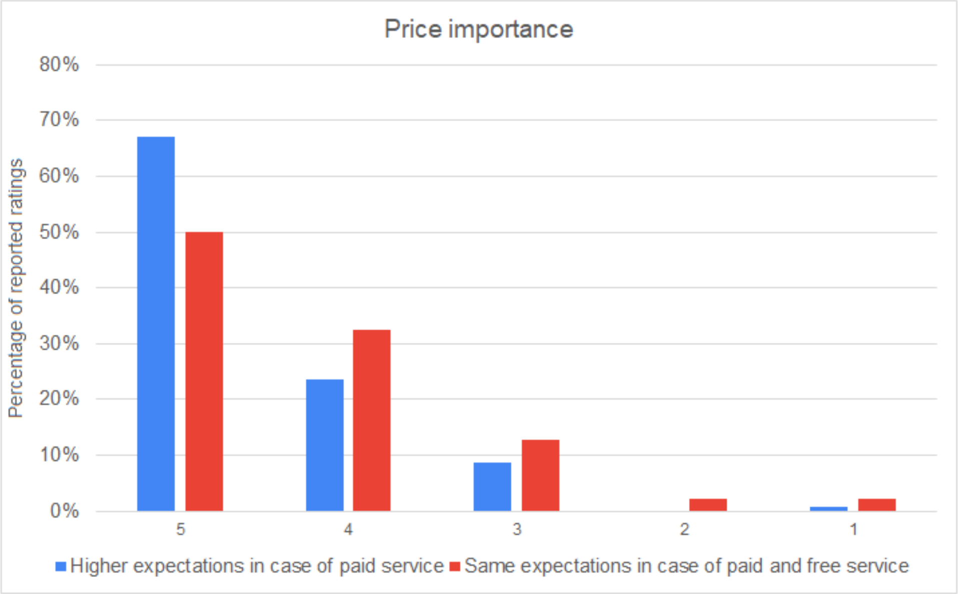


Figure 4.5: Percentage of user ratings for influence factor Price between users with a higher level of expectation for paid service and users with the same level of expectation for paid and free service.

To investigate attitudes towards price importance between groups we compared their ratings. Participants belonging to the group with higher expectations in case of a paid service in 67% consider price a very important factor, while in the group with the same level of expectations for paid and free service 50% consider price important (Figure 4.5).

We compare our obtained results to other related studies, and find that results coincide in certain aspects and showed similar results. Husić *et al.* identified in a conducted survey involving 140 participants seven most influential factors in case of WebRTC video calls: audio quality, image quality, quality of service, service price, loss of video frames, ease of use, and procedure of accessing web environment [96]. Based on this classification, García *et al.* proposed the following key performance indicators for QoE estimation: call establishment time, end-to-end delay, perceived audio, video, and audiovisual quality [97].

The results of the investigation analysis have shown two relevant areas impacting QoE, quality of real-time media (depending on application and network domain) from user perspective (audio-video synchronization, perceptible video delay, longer freezes, perceptible audio delay, reduced speech intelligibility, interrupted interaction), and application management (installation complexity, ease of use, security in terms of privacy, low battery consumption, and price).

Quality of real-time media

The ability to interact without interruptions can be difficult in face to face communication. Participants in video mediated communication can be severely affected by transmission delays, where comprehension can be distorted by mutual silence or double talk [139]. Speech intelligibility is a measure of the effectiveness of speech communication. Speech intelligibility is usually defined as the percentage of speech units (syllables, words or sentences) correctly perceived by listeners. Reduced intelligibility occurs due to the nature of the spoken material (unfamiliarity with the speaker, possible abnormal speech features or unfamiliarity with the conversation topic) and the context of transmission [140]. Speech intelligibility depends on audio bandwidth, channel impairments, input (microphone) and output (speaker) of end user devices characteristics and its placement in relation to the speaker/listener, acoustical properties of the room, sound pressure level, and background noise level [141]. If the cause of poor intelligibility does not lie in human characteristics, yet involves system components, there is a possibility to isolate the cause of the reduced quality and prevent further degradation.

Synchronization of audio and video should *provide the feeling that the speaking motion of the displayed person is synchronized with that person's voice* [85]. AV synchronization is skewed due to the impaired temporal synchronization between media and may be the result of increased transmission delays with significant delay variances and the fact that audio can be processed faster than the video (especially with high quality) [142]. For that reason, it is a more common situation that the audio stream arrives before video. In video calls, lip synchronization is considered important since it presents the natural relation between image and audio of the voice for the viewer/listener. Thus, in case of interactive communications, the goal is to minimize the delay between visual and voice. In WebRTC, the synchronization process is handled on the receiver side. When capturing the media (audio and video are handled separately), timestamps are assigned to raw media which are then encoded and sent to the receiver over the network. On the receiver side, sent packets are collected in a jitter buffer, which serves to align audio and video streams and order the packets to achieve lip synchronization.

Common types of delay are handling delay, transmission, and queuing delay. In terms of interactive conversation, audio delay can impact video call quality when participants start to double talk and interrupt each other. In situations where only one user is talking and the rest are listeners, echo can be noticed more easily [17]. Video delay can have significant impact if the task or the objective of the telemeeting relies on the visuals. In such scenarios, where visuals play an important role, longer video freezes can have significant impact on the overall quality. Lost packets lead to freeze frames which might take long to recover, and in worst case scenarios can lead to complete loss of the video. Thus, the main challenge is to keep the packet loss rate low for a longer time period [143]. In a mobile environment, where end user devices have limited capabilities, longer video freezes can cause high CPU utilization as well.

Application management

Application installation can be highly demanding, especially if prerequisite components are needed or a specific order of installation should be followed. Some components can be difficult to deploy, requiring integration with existing technologies. Applications based on WebRTC technology are plug-in free (no need to install third-party components) relying on the framework already supported within the browser. An application is loaded within the supplied browser, providing the possibility to set up a video call, meaning that additional installation is not required.

Ease of use can shape an end user's attitude toward using a service, and the intentions to use the technology. In WebRTC scenarios, to initiate the call user registration may be requested. A link name for entering a virtual meeting room must be shared among participants, and sharing could take place over some other services such as email, Facebook or Twitter. However, in cases where not all participants are able to use offered services it is good to provide a link which can easily be remembered [39].

Data protection and online privacy have the objective to guard sensitive data from unauthorized access to the data, cyber attacks, and accidental or deliberate data loss [144]. Data privacy is focused on rules considering proper handling (consent), processing, storage and usage of personal information. Data protection regulation and directives are focused on the confidentiality, integrity and availability of information.

Battery consumption is an important aspect of the mobile user experience [145]. Thus, in a mobile environment, an application must carefully utilize display (high-resolution touchscreen), camera, and radio resources.

Price is a one of the influence factor that impacts consumer behavior and purchase decisions in all aspects of living, so there is no difference in the IT industry as well. The pricing strategy deployed by providers to price their products or services will be based on the ability of users to pay, market conditions, competition, and input costs. All of those factors serve as a guidance for control and affording choice and have to be considered in application design and development.

The majority of identified QoE IFs come from the application and network domain, and serve as an input for defining a methodology to derive an adaptation strategy based on the video encoding parameters. Price, data privacy, installation complexity, and easy of use are the influence factors on the application management level, which do not interfere with the video encoding parameters. Thus, in further research those factors will be out of focus.

To quantify QoE based on the perceived media quality, the focus should be on: speech intelligibility, audio-video synchronization, longer video freezes, perceptible audio delay, image blurriness, image sharpness, and perceptible video delay. The cause of all listed impairments can lie in end user device processing capabilities, network disturbances or choice of inadequate coding parameters values with respect to the content/service type and available resources.

Discussed factors that should be considered can be controlled, at least to some extent from an application point of view, by video encoding parameters, impacting the system as a whole. On application level, it is possible to adapt video quality level (e.g., resolution, bitrate, and frame rate) to avoid CPU overload which can lead to congestion, and to save bandwidth needed for transmission and retain acceptable QoE (Figure 4.6).

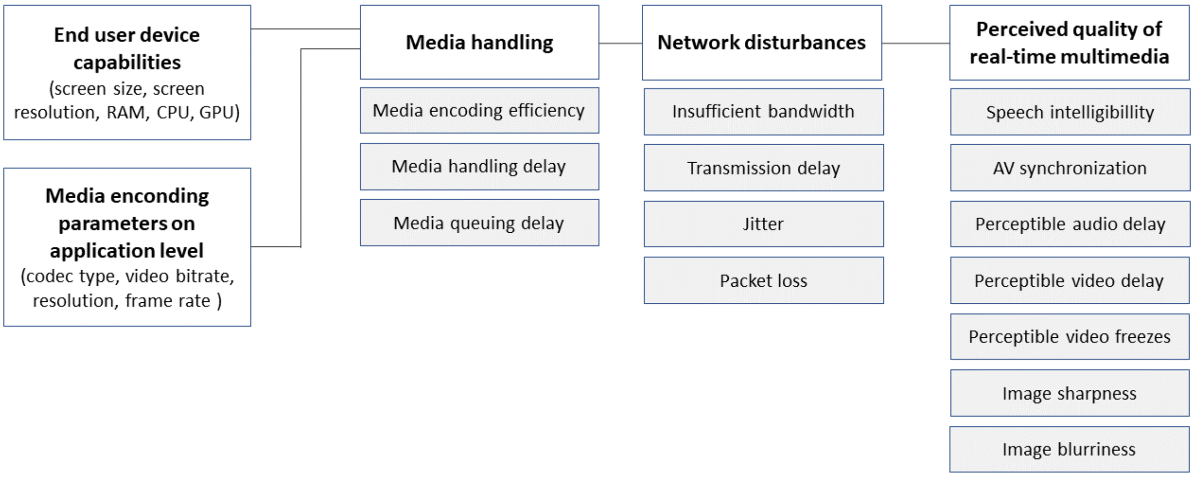

End user device capabilities
(screen size, screen resolution, RAM, CPU, GPU)
Media enconding parameters on application level
(codec type, video bitrate, resolution, frame rate)
Media handling
Media encoding efficiency
Media handling delay
Media queuing delay
Network disturbances
Insufficient bandwidth
Transmission delay
Jitter
Packet loss
Perceived quality of real-time multimedia
Speech intelligibillity
AV synchronization
Perceptible audio delay
Perceptible video delay
Perceptible video freezes
Image sharpness
Image blurriness

Figure 4.6: Example of selected parameters that impact perceived quality of real-time media in a mobile context in the end to end pipeline.

4.5 Impact of human influence factors on opinions

Part of the challenges in QoE assessment and modeling can be attributed to the human diversity. We explored user expectation differences in terms of age, gender and education level, taking into account number of reported ratings per specific group. In our case, we assume that if the difference between the two groups is 10% or more, then group impact is present. If it is less than 5%, then there was little, if any, group impact. We can observe that the younger the user group, the impact of influence factors is considered more often as “very important”. Figure 4.7 shows that there are generational differences in reported ratings of the impact of influence factors. The likelihood of given ratings does not differ significantly between millennials and those 35 years and younger, as well as participants between 46 years and older and older adults. According to the reported ratings, young adults aged between 18-25 and 26-36 rated the impact of influence factors with 5 in 43.44% and 40.84% of cases, respectively. On the other hand, middle-aged adults 36-45, 46-55 and older adults >55 rated the impact “very important” or with a “great extent” with the share of 33.05%, 28.39% and 28.67%, respectively (Table 4.6).

Results also show that female participants (39.4%) are 1.52 times more likely to consider impact factors “very important” than male participants (25.81%). In contrast, the portion of ratings 4 reported by female participants (33.28%) fell by 6.06%, meaning that male participants (41.93%) were 1.26 times more likely to rate impact factors as “important” (Figure 4.8). Rating 1 was reported by 6.46% male and 3.53% female participants (Table 4.6). Results suggest that there are some gender-related differences when perceiving factor importance.

In terms of application frequency usage, we categorized participants per user type as: **very frequent user**, uses application on daily basis, **frequent user**, uses 2 to 3 times per week, **occasional user**, uses application 4 to 7 times per month while the **light user** category includes users that use video conferencing/telemeeting applications 3 or less times per month. Of the

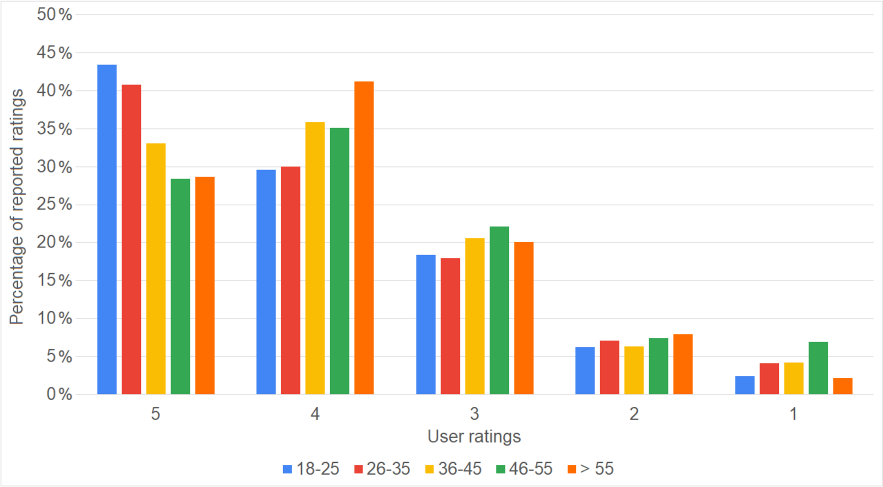


Figure 4.7: Percentage of reported ratings across all questions with respect to age.

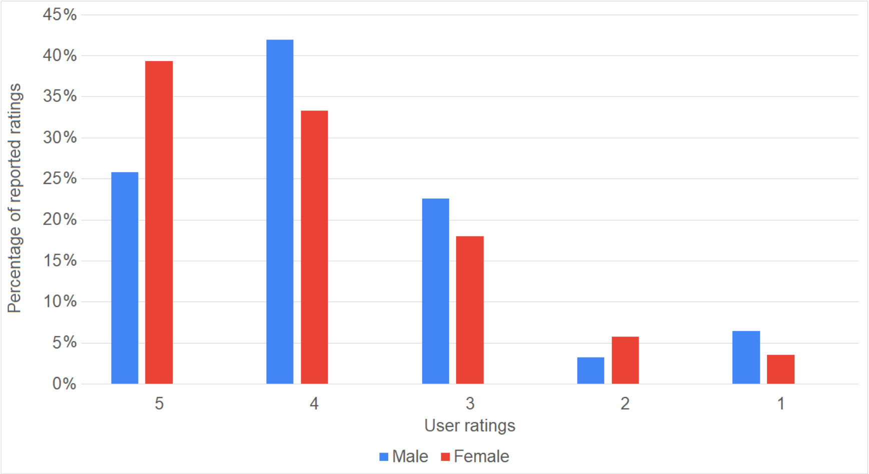


Figure 4.8: Percentage of reported ratings across all questions with respect to gender.

Table 4.6: Percentage of reported ratings distributed between participant's age and gender.

User	5	4	3	2	1
Age					
18-25	43.44%	29.57%	18.39%	6.24%	2.36%
26-35	40.84%	29.98%	17.98%	7.11%	4.09%
36-45	33.05%	35.89%	20.59%	6.29%	4.18%
46-55	28.39%	35.16%	22.15%	7.42%	6.88%
> 55	28.67%	41.22%	20.07%	7.88%	2.15%
Gender					
Male	25.81%	41.93%	22.58%	3.22%	6.46%
Female	39.34%	33.28%	18.05%	5.80%	3.53%
User type					
Very frequent user	42.55%	33.33%	16.2%	5.79%	2.13%
Frequent user	36.45%	33.06%	21.13%	6.13%	3.23%
Occasional user	33.66%	35.34%	19.28%	7.57%	4.15%
Light user	33.31%	33.98%	20.81%	6.79%	5.11%
Education level					
Higher University degree (PhD)	34.79%	34.77%	20.89%	6.02%	3.53%
University degree (bachelor or masters)	35.01%	29.75%	17.92%	9.79%	7.53%
High school degree	38.27%	32.38%	16.99%	7.45%	4.91%

participants polled, 42.55% from the **very frequent** user category reported the selected influence factors as "very important" (Figure 4.9). **Frequent** users were likely to rate an impact factor as "very important" with 36.45% answering, while both **occasional** and **light** users were likely to rate the same, at 33.66% and 33.31%, respectively (Table 4.6). Based on the results we can conclude that the difference between user type is not significant.

Focusing on the attitudes toward factor importance between participants with different education levels, we did not find significant differences (Figure 4.10). Namely, users with a high school diploma (38.27%) are more likely to consider a factor as "very important" than users with a bachelor and masters (35.01%) or users with a PhD (34.79%) (Table 4.6).

Additionally, statistical relationships between mean ratings of impact factor importance and age, gender, and education level were reported in table 4.7. We also included relationships be-

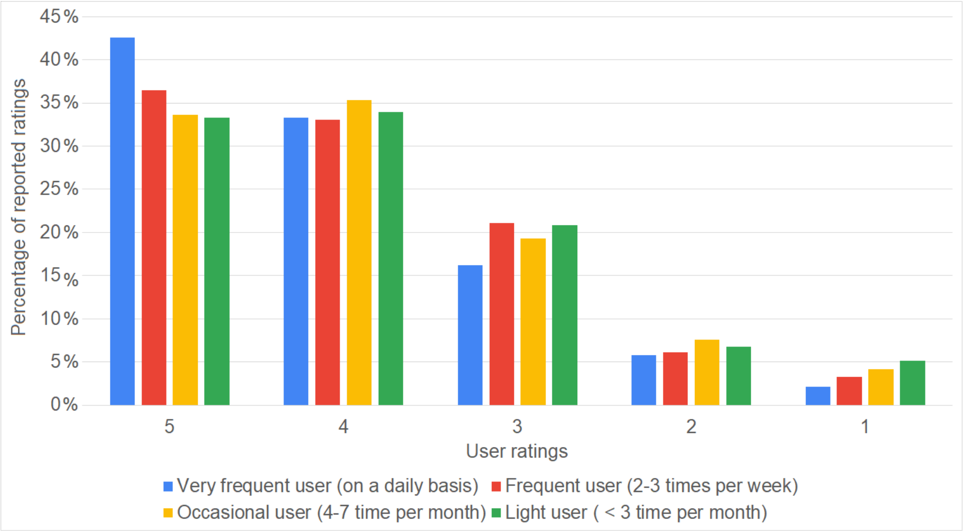


Figure 4.9: Percentage of reported ratings across all questions with respect to frequency of application usage.

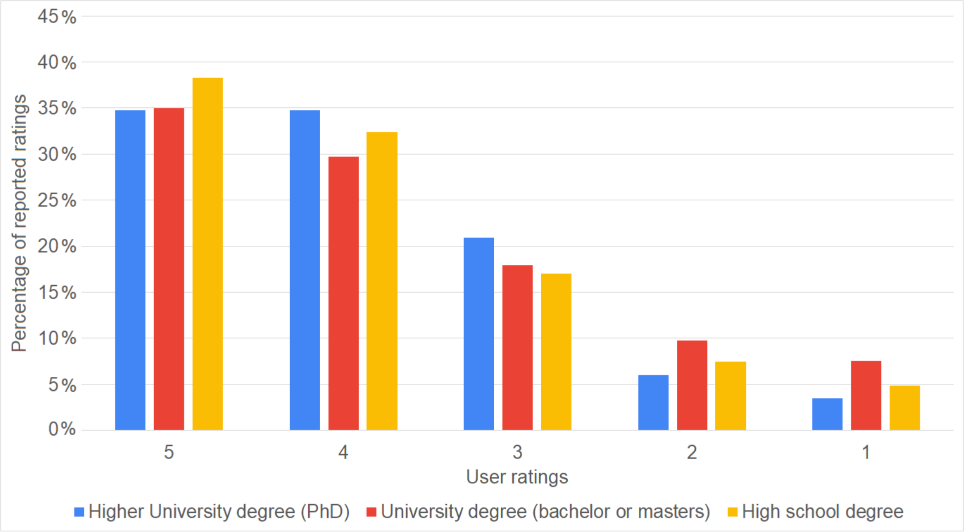


Figure 4.10: Percentage of reported ratings across all questions with respect to education level.

Table 4.7: Correlations between ratings and human IFs (age, gender, education level, and user type (** p-value < 0.01).

Impact Factor Importance	Age	Gender	Education level	User type
mean rating per all factors	-0.368**	0.187**	-0.068	-0.102
factors rated very influential (5)	-0.425**	0.185**	-0.051	-0.097

tween rating 5 of impact factor importance and age, gender, and education level. To measure the strength of the relationship, we calculated Pearson's correlation coefficients. The data shows moderate negative correlation between perceived IFs importance and age, positive weak correlation between perceived IFs importance and gender, while no correlation was found between education level or user type and perceived importance of IFs.

4.6 Chapter summary

This chapter summarizes the results of a survey conducted among 272 participants designed to investigate users' opinions and expectations related to audiovisual telemeetings on mobile devices in the leisure/private context. Results have verified the relevant conclusions from existing literature in terms of perceived quality, influence factors, and user expectations related to multiparty audiovisual telemeetings. Based on the conducted survey and our own user studies, we identified key system-, context-, and human-related factors and corresponding QoE features/dimensions (Figure 4.11).

Based on user ratings, we selected the most influential factors in descending order: speech intelligibility, audio-video synchronization, longer freezes, perceptible audio delay, low battery consumption, image blurriness, price, security in terms of privacy, ease of use, perceptible video delay, uninterrupted interaction and installation complexity.

Additionally, we identified age and gender as the most influential human factors in terms of expectations, and direct perception, while the most influential context IFs are related to the multiparty setup (mobility, number of participants, site distribution and use case). The listed factors and features can serve as input for QoE modeling in the case of a multiparty audiovisual telemeetings on mobile devices. Having addressed influence factors belonging to all three groups (human, system, context), in following chapters we focus on objective video quality in terms of video encoding parameters, and the impact on subjective user ratings and objective quality metrics.

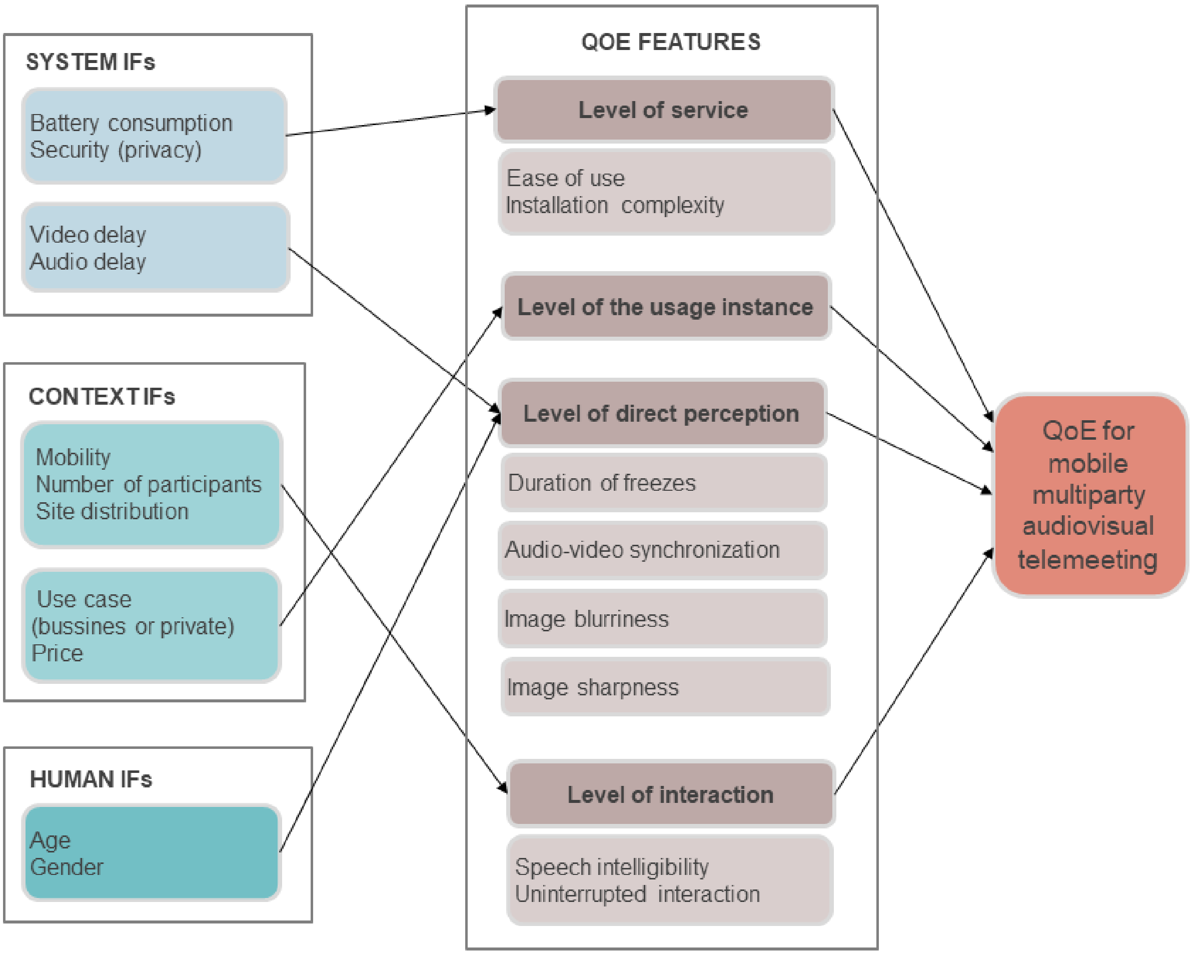


Figure 4.11: Key system-, context-, and human-related influence factors and corresponding features for multiparty audiovisual telemeetings on mobile devices as derived based on the survey results, conducted studies, and standards. Please note that numerous underlying factors impact QoE features and overall QoE (such as coding configurations and network impairments).

Chapter 5

Impact of system factors on QoE for multiparty audiovisual telemeetings on mobile devices

Deploying multiparty video communication solutions on smartphones calls for the need to optimize video encoding parameters due to limited device processing power and dynamic wireless network conditions. Given the mobile device context and corresponding screen sizes, the question arises as to which video quality levels should be maintained during a call so as to achieve acceptable QoE. In other words, increasing video quality beyond a certain threshold will likely not contribute to user perceivable QoE improvement. In cases of variable and limited system and network resource availability, video encoding adaptation strategies may be deployed to downsize traffic by adapting parameters such as bitrate, resolution, and frame rate, so as to optimize end user QoE.

Our main research focus has thus been geared towards deriving QoE-driven service adaptation strategies, based on the adjustment of video encoding parameters in accordance with available resources. Given the wide range of potential test conditions, in this and the following chapter we report on six subjective user studies we have conducted over the course of four years (2015-2018, as summarized in Table 1.1), and highlight the main findings of each study. The studies were aimed to collect a large number of subjective ratings under various conditions, and obtain insights into both session and stream quality. Following the collection of subjective ratings, we further aimed to investigate the potential of utilizing objective video metrics to infer subjectively perceived quality (studies US5 and US6).

Studies were based on the investigation of the impact of different video encoding parameters and smartphone capabilities on QoE (Table 5.1). We highlight here the general methods used, referring to all conducted studies, while specific methods applied in each study are described in more detail in the corresponding subsections of Chapters 5 and 6. The test setup included three-

Table 5.1: Characteristics and capabilities of smartphones used over the course of reported studies.

Parameter	Samsung SIII	Samsung S5	LG G3	Samsung S6	Samsung S7
Chipset	Exynos 4412 Quad	Qualcomm MSM8974AC Snapdragon 801	Qualcomm MSM8974AC Snapdragon 801	Exynos 7420 Octa	Exynos 8890 Octa
CPU	Quad-core 1.4 GHz Cortex-A9	Quad-core 2.5 GHz Krait 400	Quad-core 2.5 GHz Krait 400	Octa-core (4x2.1 GHz Cortex-A57 4x1.5 GHz Cortex-A53)	Octa-core (4x2.3 GHz Mongoose 4x1.6 GHz Cortex-A53)
GPU	Mali-400MP4	Adreno 330	Adreno 330	Mali-T760MP8	Mali-T880 MP12
RAM	1 GB	2 GB	2 GB	3 GB	4 GB
Display size	4.8"	5.1"	5.5"	5.1"	5.1"
Display resolution	720 x 1280 px	1080 x 1920 px	1440 x 2560 px	1440 x 2560 px	1440 x 2560 px

party symmetric (user studies US2, US3, US4, US5, and US6) and asymmetric (user study US1) conditions in both natural home (user studies US1, US2, and US3) and laboratory environments (user studies US4, US5, and US6). Communication flows were realized via both the public Internet, and in a controlled local area network. In all cases, the audiovisual telemeetings were realized via the WebRTC paradigm over UDP [146], as such a setup enabled us to configure encoding parameters and access session related performance statistics (via the *webrtc-internals* tool).

Selected participants ages ranged from 20 to 65, and all were non-experts in the AV field. All participants had good hearing and viewing abilities (some with corrected vision - glasses or lenses). It is important to highlight that all measurements were conducted in a **leisure context** between three acquaintances, ensuring a smooth and continuous flow of conversation. The conversations were all conducted using the Croatian language, as this was the native language for all participants. Participants did not use written materials and they were instructed to use natural conversation without any predefined task, trying to retain their attention on the mobile device. For all studies, participants were located in separated rooms, one person per room, with similar acoustics and background noise characteristics as well as video backgrounds. We performed measurements both during daylight and with artificial lights, avoiding direct light sources on the participants and cameras.

The placement of the camera and microphone on the smartphone in relation to the participants was arbitrary. In all of our tests, participants were free to hold the smartphone in their hand or place it on a stand provided to them at the viewing distance and position they preferred.

In all user studies, at the beginning of each test session, a preliminary test was carried out aimed to familiarize participants with the task and assessment questionnaire, and to make sure they felt comfortable during the evaluation. Preliminary results were not taken into account.

To prevent participant fatigue, we adhered to relevant standards, which state that the total number of tests must be reasonable and limited [11]. The total time for testing should be

Table 5.2: An overview of conducted subjective QoE studies.

User Studies	Participant, MIN/MAX/AVG age	End user device	Manipulated parameters	Collected measures
US1, 2015, [39]	18 males, 12 females, 29/65/35	Samsung S5, Samsung S3, LG G3	Device capabilities	Subjective ratings
US2, 2016, [40]	14 males, 13 females, 32/65/38	3 x Samsung S6	Video resolution, bitrate	Subjective ratings
US3, 2017, [5]	16 males, 14 females, 1 fixed user per test group, 33/49/40	3 x Samsung S6	Video resolution, bitrate, frame rate, packet loss	Subjective ratings
US4, 2018, [41]	21 males, 6 females, 20/29/21	3 x Samsung S6	Video resolution, bitrate, frame rate	Subjective ratings
US5, 2018, [42]	7 males, 20 females, 20/25/22	3 x Samsung S7	Video resolution, bitrate, frame rate	Subjective ratings and objective video quality (blurriness and blockiness)
US6, 2018, [43]	16 males, 11 females, 23/23/28	3 x Samsung S7	Video resolution, bitrate, frame rate	Subjective ratings and objective video quality (blurriness and blockiness)

balanced with respect to the time spent engaging in the service per test condition. Thus, to prevent fatigue, experiments were limited to a maximum one hour duration, and participants were given 5 minute breaks between each test condition.

In our user studies, “subjective quality” refers to user ratings obtained using an ACR scale: 1 “Bad”, 2 “Poor”, 3 “Fair”, 4 “Good”, 5 “Excellent” (used to calculate MOS), depending on the audio and video under evaluation and evaluation context (leisure, multiparty, mobile end user device). After the completion of each test condition, participants were asked to rate overall quality, audio quality, video quality, and AV synchronization using the 5-pt. ACR scale. Even though participants were asked to rate audio quality and synchronization, in-depth insights on types of distortions were not identified. We only focused on the video quality and visual impairments.

Call initiation was not in the focus of the studies, so the video call was established by the test administrator if needed. In general, a user initiating the video call creates an online virtual room via a web application to start a WebRTC session. Other users are invited to access the room through their web browser and generated URL. Before joining the room, each user must grant the service permission to access their camera and microphone.

Table 5.2 gives a brief overview and summarizes differences between the user studies. Each study is then described in detail in the following sections as follows: user study US1 (Section 5.1) investigates the impact of different end user device configurations on QoE, while user studies US2 (Section 5.2), US5 (Section 5.3) and US6 (Section 5.4.2) investigate the impact of different video encoding parameters on the perceived video quality. Additionally, in studies US5 and US6 we included the analysis of objective video quality metrics (blurriness and blockiness) in order to establish the relationship between perceived quality and objective video quality.

5.1 User study US1 - Impact of end user device capabilities

Our initial research focused on studying the impact of different smartphone configurations (differing in terms of CPU, display size, and resolution) on QoE (results published in [39]). Tests were run using available WebRTC-based conferencing applications available on the market at the time of the study.

5.1.1 Methodology

Tests involved interactive three-party audiovisual conversations in a natural environment with telemeeting set-up on mobile phones and laptops over a WLAN and commercial network. Experiments were conducted involving three different general set-ups, namely:

- all three participants in the group had the same smartphone configuration,
- each participant in the group had a different smartphone configuration, and
- all three participants in the group used different laptops.

The three-party video conference was set up using two different schemes: (1) using WebRTC applications running on the Internet, and (2) using the Kurento Media Server (KMS) installed in a local network (Figure 5.1).

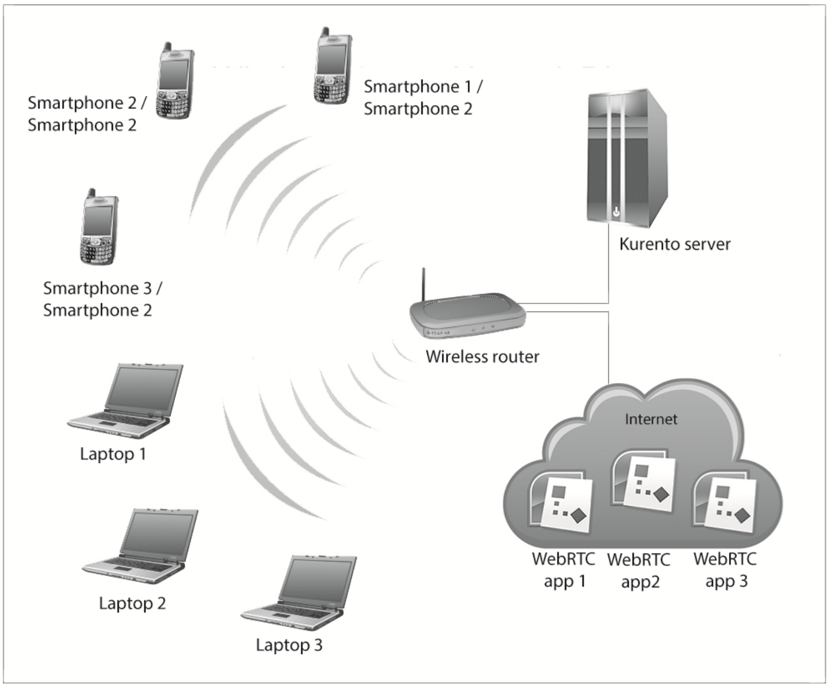


Figure 5.1: System set-up over LAN and Internet (user study US1).

Kurento[1] is an open source WebRTC media server and offers a set of client APIs intended for development of advanced video applications. For our testbed setup we used the Kurento 5.0.5. Media Server installed on a laptop with Intel Core i5 Processor 3230M, 2.6 GHz, 8 GB

[1] http://www.kurento.org/

RAM and Ubuntu 14.04 LTS. The LAN connection between end user devices and the media server is Wi-Fi 802.11b, on port 8080.

There were many vendors offering WebRTC video communication services, but at the time this study was conducted, only few of them provided free, no login, no installs multi-party video chat support. Tests were run using three publicly available web applications: talky.io, appear.in, and vline. All three applications used the same topology (P2P) and methodology of creating, starting, and terminating conversations.

During the video conversation, participants did not use available additional functionalities, hence they are not tested as a part of the service evaluation and consequently did not affect QoE in our experimental setup. However, user opinions regarding the need for additional functionalities were investigated with paper questionnaires after the subjective assessment of the WebRTC video service.

The first WebRTC application Talky[2] (WebRTC App 1) allowed to set up a video conversation with up to 5 participants. The URL to start a session was simple to remember. The application has easy access to additional functionalities including screen sharing, mute, hold video or lock the room.

The second application was Appear.in (currently under the name Whereby)[3] (WebRTC App 2), which enables setting up a multi-party conversation with up to 8 participants. The URL was easy to remember. Additional functionalities included text messaging, mute, disabling camera, lock or leave the room.

An important consideration for multiparty video conversation service deployment is the perceived ease-of-use. To join a conversation with the third application vLine[4] (WebRTC App 3), a participant can use another service. A participant which creates a room must enter a name, then share a link, which is not intuitive and easy to remember. The link can be shared via other services such as email, Facebook, or Twitter, which we did not use during our tests.

To explore the effects of different end user device configurations, we used three different smartphone (Samsung SIII, Samsung S5, and LG G3) and three different laptop configurations, as specified in Table 5.1 and Table 5.3.

Tests involved the following set-ups: (1) all three participants in the group had the same smartphone configuration, (2) each participant in the group had a different smartphone configuration. Tests were conducted in a natural environment on mobile phones over both a Wi-Fi and commercial mobile network. The test schedule for 12 conditions based on 3 minutes per condition length is shown in Table 5.4.

A real-time communication service typically delivers either text, audio, graphics, video and data, or some combination of the aforementioned media types. Therefore, the second part of

[2] https://talky.io/

[3] https://whereby.com/

[4] https://vline.com/

Table 5.3: Laptop characteristics.

Parameter	Laptop 1	Laptop 2	Laptop 3
CPU	Intel Core i7 2670Q Processor, 2.2GHz	Intel Core i7 4700NQ Processor, 2.4GHz	Intel Core i7 4710HQ Processor, 2.5GHz
RAM	6 GB	16 GB	16 GB
Display size	15.6"	15.6"	17.3"
Display resolution	768x1366 px	1080x1920 px	1080x1920 px
Camera	1.0 MP (1280x720)	2.0 MP (1920x1080)	1.0 MP (1280x720)
OS	Windows 7	Windows 8.1	Windows 8.1
Web browser	Chrome 41.0.2272.89 m	Chrome 41.0.2272.89 m	Chrome 41.0.2272.89 m

our questionnaire was intended to explore the possibility of enhancing the user experience with additional functionalities. Participants were asked: *"Would the listed functionality enhance QoE?"* (Yes/No). The tested WebRTC applications have some functionalities already included on Web pages, but in this experiment they were not utilized, given that the three-party video communication itself presented a great load for the smartphones.

Tests were organized as a mix of between-subject and within-subject design, as shown in Table 5.4. Overall 30 participants took part in the study and were divided into ten fixed groups with 3 members each. The groups were mixed with respect to gender, with a total of 18 male and 12 female participants taking part in the studies. The average age was 35 years, while the youngest participant was 29 and the oldest 65 years old.

Given that the testing did not require a high sensitivity to different test conditions, it was not necessary for participants to have had previous experience with multiparty conversational systems. The selected participants had no special knowledge of audio/video technology, no experience with subjective test methodologies and had not participated previously in subjective assessments, nor were they technical experts regarding the equipment and services to be tested. However, all participants reported using smartphones on a daily basis. Participants were comprised of volunteers, and all have normal or corrected vision and normal hearing.

5.1.2 Results

Statistical analysis was employed to interpret experiment results. A one-way Analysis of Variance (ANOVA) was used to find significant differences between the smartphone configurations with respect to subjective quality ratings for different WebRTC applications. The results of the 1-way ANOVA confirm that the difference between smartphone 1 and smartphone 2 or smartphone 3 is significant at 95% significance level in the case of Web Application 2 for overall and audiovisual quality as well as interaction reduction (shown in Table 5.5). To identify if variations between smartphone 2 and smartphone 3 are significant, a t-test two-sample was applied.

Results showed that configuration smartphone 2 and 3 when used with WebRTC Application

Table 5.4: Test schedule used in user study US1.

Test case (TC)	Participant	End user device	Application (used in test case)	Network connection	Time[min]
Instructions					10
TC1, TC2, TC3	A	Smartphone 1	WebRTCapp 1 (TC1)	Wi-Fi, Internet	3
	B	Smartphone 2	WebRTCapp 2 (TC2)		
	C	Smartphone 3	WebRTCapp 3 (TC3)		
Questionnaire subjective assessment					
Break					5
TC4, TC5, TC6	A	Smartphone 2	WebRTCapp 1 (TC4)	Wi-Fi, Internet	3
	B	Smartphone 2	WebRTCapp 2 (TC5)		
	C	Smartphone 2	WebRTCapp 3 (TC6)		
Questionnaire subjective assessment					
Break					5
TC7	A	Smartphone 1	Kurento (TC7)	WLAN	3
	B	Smartphone 2			
	C	Smartphone 3			
Questionnaire subjective assessment					
Break					5
TC8	A	Smartphone 2	Kurento (TC8)	WLAN	3
	B	Smartphone 2			
	C	Smartphone 2			
Questionnaire subjective assessment					
Break					5
TC9	A	Laptop 1	Kurento (TC9)	WLAN	3
	B	Laptop 2			
	C	Laptop 3			
Questionnaire subjective assessment					
Break					5
TC10, TC11, TC12	A	Laptop 1	WebRTCapp 1 (TC10)	Wi-Fi, Internet	3
	B	Laptop 2	WebRTCapp 2 (TC11)		
	C	Laptop 3	WebRTCapp 3 (TC12)		
Questionnaire subjective assessment					

Table 5.5: ANOVA analysis for smartphone difference.

Source of variation	SS	MS	F	P-value	F crit
Overall quality					
WebRTC app 1	0.1	0	0.26	0.76	3.35
WebRTC app 2	7.4	3.7	27.88	3.945E-3	3.35
WebRTC app 3	0.1	0	0.5	0.61	3.35
Audiovisual quality					
WebRTC app 1	0.2	0.1	0.58	0.56	3.35
WebRTC app 2	2.4	1.2	9.21	0.11	3.35
WebRTC app 3	0.2	0.1	1.08	0.35	3.35
Interactivity reduction					
WebRTC app 1	0.2	0.1	1.08	0.35	3.35
WebRTC app 2	2.86	1.43	9	0.06	3.35
WebRTC app 3	0.06	0.03	0.5	0.61	3.35

Table 5.6: ANOVA analysis for webrtc application difference on smartphone 2.

Source of variation	SS	MS	F	P-value	F crit
Overall quality	51.26	25.63	126.47	1.79E-26	3.1
Audiovisual quality	66.15	33.07	132.82	3.63E-27	3.1
Interactivity reduction	85.48	42.74	181.69	8.66E-32	3.1

2 for overall and audiovisual quality as well as interaction reduction can be considered as equal. In all test results, it should be noted that potential order effects may have occurred due to the experimental design (order of test scenarios). In the case of WebRTC Applications 1 and 3, contrary to expected, there was no evidence that significant differences between smartphones exist. We therefore looked to establish whether or not there were significant differences in quality ratings between the WebRTC applications themselves. To test the difference between WebRTC applications, we analyzed the scenarios where each participant within each group used smartphone 2 (in order to keep the device factor constant). Although all applications are based on fully meshed peers, presented ANOVA results confirm that the difference between WebRTC applications is significant for overall and audiovisual quality as well as interaction reduction ratings, as shown in Table 5.6.

Statistics were performed using a 95% significance level. To observe the difference between WebRTC applications, a t-test was applied. Results again reveal that WebRTC applications 1 and 3 used with smartphone 2 can be considered as equal, while results between applications

Table 5.7: ANOVA analysis for WebRTC application difference on laptops.

Source of variation	SS	MS	F	P-value	F crit
Overall quality	1.75	0.87	2.87	0.06	3.1
Audiovisual quality	1.26	0.26	2.38	0.09	3.1
Interactivity reduction	0.46	0.23	0.81	0.44	3.1

Table 5.8: Average quality ratings combining ratings collected across all smartphones (1, 2, and 3).

WebRTC application	Overall quality	Audiovisual quality	Interactivity
WebRTC app 1	1.16	1.13	1.06
WebRTC app 2	2.93	2.9	2.8
WebRTC app 3	1.03	1.16	1.06

1 and 2 as well as between applications 2 and 3 suggest significant difference. We note that explicit monitoring of smartphone CPU and memory usage was not observed. Another limitation of the study was that we were not able to detect the exact video resolution and codecs used. WebRTC applications typically use VP8 and Opus (which we detected in the case of users using laptops). However unadjusted payload and video resolutions unnecessarily too high for mobile applications can have a significant impact on QoE. Hence, our subsequent research aimed to address the effects of different video resolutions on QoE. To further explore WebRTC application capabilities while eliminating the mobile device impact on QoE ratings we repeated test procedures using laptops as end user devices. ANOVA results showed that there is no significant difference between WebRTC applications running on laptops at 95% significance level and they can be considered as equal for overall and audiovisual quality as well as interaction reduction (results shown in Table 5.7).

In each test scenario participants reported distortion detection as delay and freezing. Experiments based on the different smartphones had unacceptable quality for each WebRTC application. For Laptop sessions and smartphone 2 sessions participants evaluated quality as acceptable. Comparing average results for smartphone and laptop session findings have shown the impact of different end user devices on QoE. WebRTC application 2 showed the best performance both in the smartphone and laptop scenario with an average overall quality rating of 2.93 in different smartphones sessions (Table 5.8), 3.36 for smartphone 2 sessions (Table 5.9), and 3.76 for laptop sessions (Table 5.10).

Laptop sessions have shown that there are no significant differences between chosen WebRTC applications, likely due to the fact that the laptop devices were equipped with enough hardware resources to run all three applications. Therefore, we concluded that low ratings for WebRTC applications 1 and 3 (only in smartphone scenario), likely arise from smartphone con-

Table 5.9: Average quality ratings on smartphone 2.

WebRTC application	Overall quality	Audiovisual quality	Interactivity
WebRTC app 1	1.73	1.5	1.3
WebRTC app 2	3.36	3.4	3.36
WebRTC app 3	1.8	1.6	1.36

Table 5.10: Average quality ratings across all laptops.

WebRTC application	Overall quality	Audiovisual quality	Interactivity
WebRTC app 1	3.4	3.3	3.36
WebRTC app 2	3.76	3.56	3.5
WebRTC app 3	3.5	3.53	3.3

figuration overload. Application 2 clearly has lower processing requirements, likely due to a better adaptation strategy for mobile users. However, further research is needed to determine whether different video resolutions and codecs are being used and if so to what degree this impacts processing requirements and QoE.

A small difference between average ratings for laptop sessions and smartphone 2 sessions for overall and audiovisual quality as well as interactivity reduction showed that participants have much lower expectations for audiovisual telemeetings on smartphones then on laptops. Lower expectations in natural environments in a leisure context are reasonable since there is no specific task that must be realized and finished.

Experiments where all participants used the same smartphone to avoid variances caused by differences of smartphone configuration scored higher ratings for all questions than experiments where all participants used different smartphones with one lower end smartphone, implying that the weakest end user device has a negative impact on the QoE of each participant within a group. Figure 5.2, 5.3, and 5.4 presents average ratings per smartphone in sessions where each participant used a different smartphone. Average rating for smartphone 1 when used with WebRTC application 1 or 3 was 1.1 for overall quality, 1 for audiovisual quality and 1 for interactivity.

In the case of WebRTC application 2 the average rating for overall quality was 2.1, 2.3 for audiovisual quality and 2.2 for interactivity reduction. Measured values has shown that smartphone 1 configuration may be considered as insufficient for a three-party video conversation, while smartphone 2 or 3 may be considered as having minimum hardware configurations necessary to run the service with fair QoE.

ANOVA results clearly showed, and average quality rating results confirmed, three important findings: (1) that the smartphone configuration has a significant impact on QoE, (2) the

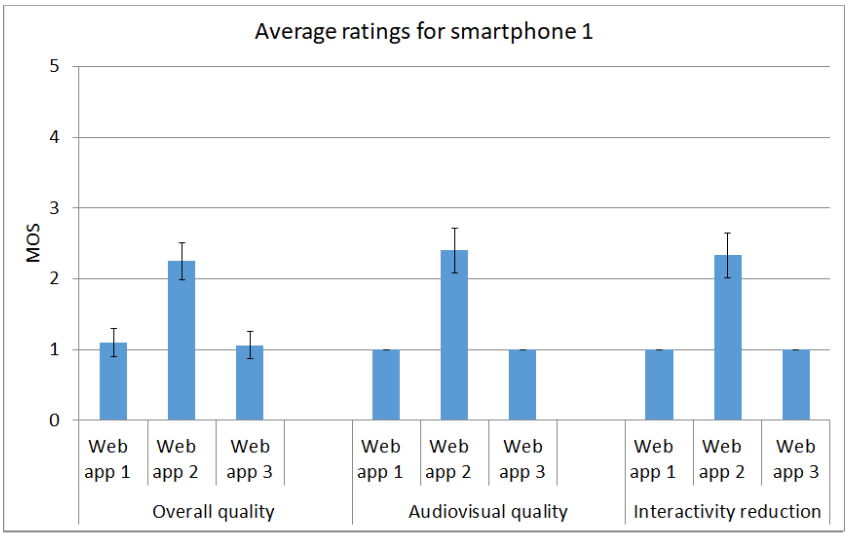


Figure 5.2: Average quality ratings for participants using smartphone 1 (95% CI shown).

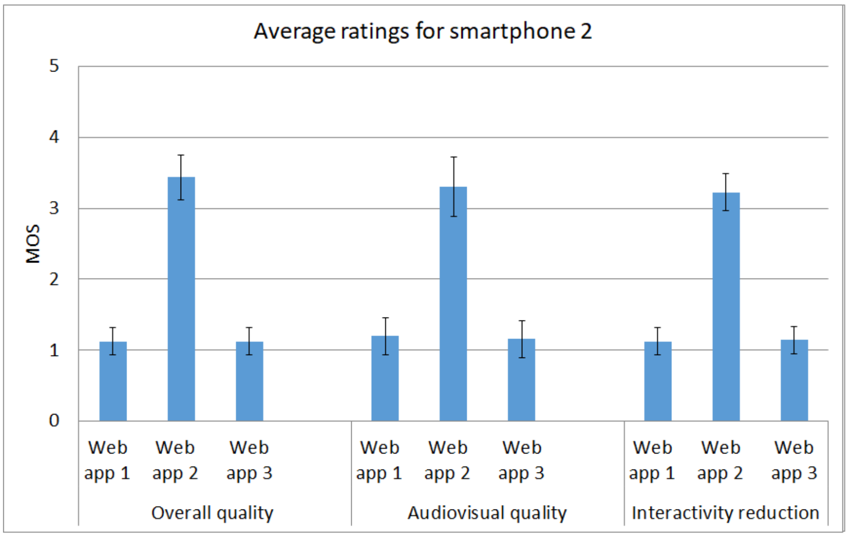


Figure 5.3: Average quality ratings for participants using smartphone 2 (95% CI shown).

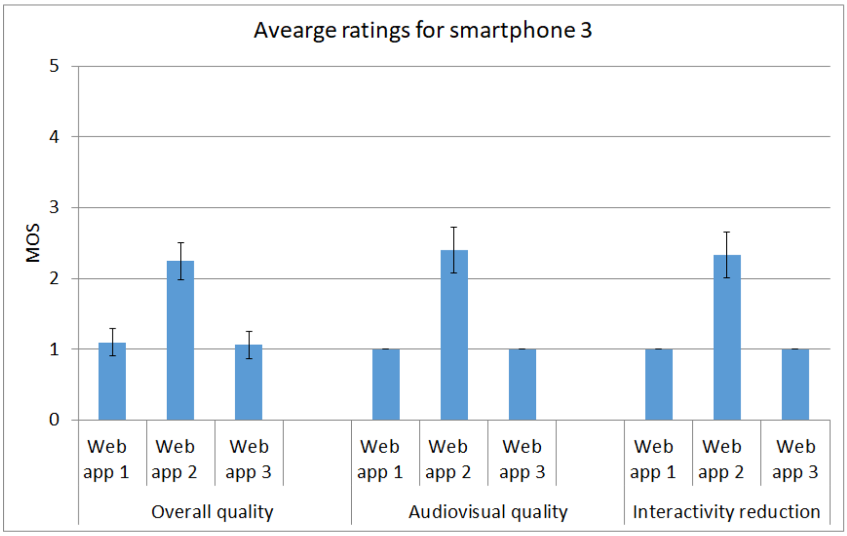


Figure 5.4: Average quality ratings for participants using smartphone 3 (95% CI shown).

Table 5.11: ANOVA analysis of overall quality, audiovisual quality, and interactivity rated on smartphone 2 and using Kurento.

Source of Variation	SS	MS	F	P-value	F crit
Overall quality	14.06	7.03	31.13	9.76E-08	3.35
Audiovisual quality	6.86	3.43	13.24	9.82E-05	3.35
Interactivity reduction	6.66	3.33	12.5	1.44E-04	3.35

Table 5.12: Average quality ratings on Kurento.

Kurento	Overall	Audiovisual	Interactivity
Smartphone 1,2 and 3	3.03	2.93	3.1
Smartphone 2	3.73	3.43	3.53
Laptop	4.66	4.63	4.56

WebRTC application for smartphones should be optimized for lower processing power, and (3) participants have greater expectations for audiovisual telemeetings on laptops then on smartphones.

The last experiment was conducted with Kurento in a local area network without any influence of impairments, such as packet loss or delay, caused by running tests via a real commercial network (Figure 5.5).

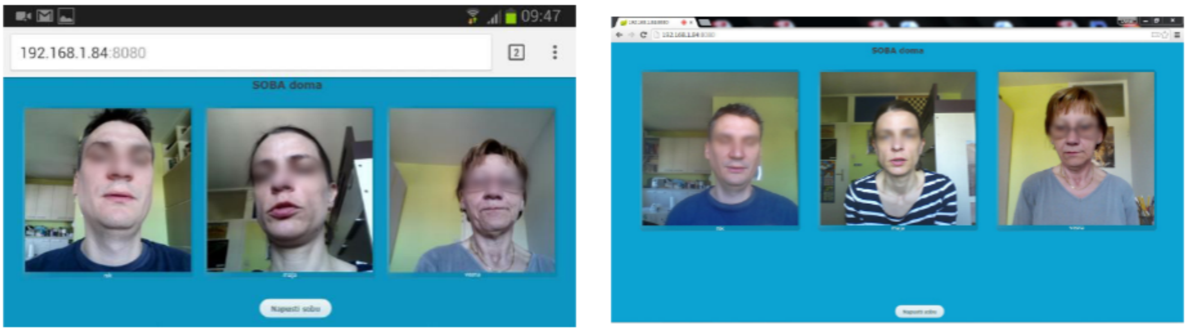


Figure 5.5: Example of video call established over Kurento and smartphone and laptop.

ANOVA results for the Kurento web application also confirm that the difference between Smartphone 1, smartphone 2 and smartphone 3 is significant (Table 5.11). In this case, a t-test also identified variations between smartphone 2 and smartphone 3 as not significant.

Experiments conducted using Kurento were assessed with higher ratings than with commercial WebRTC applications (Table 5.12). The low-end smartphone 1 scored better ratings than with commercial WebRTC applications. Understandably, the public network did have an impact on ratings.

Nevertheless, the average rating evaluated with smartphone 1 was 2.2 for overall quality, 2.4

for audiovisual quality, and 2.3 for interactivity reduction (Figure 5.6). Therefore, the smartphone 1 configuration may again be considered as inadequate for three-party video conversation. With an average rating of fair quality, Smartphones 2 or 3 may again be concluded as having minimum hardware configurations necessary for sufficiently high perceived quality.

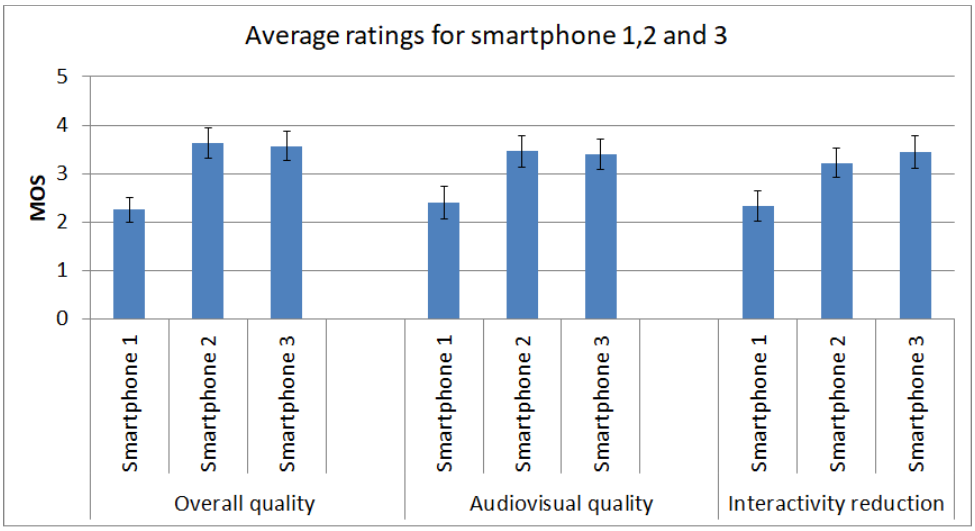


Figure 5.6: Average quality and interactivity ratings for smartphones 1, 2, 3 over Kurento (95% CI shown).

Finally, conducted experiments showed that the smartphone 1 configuration was not capable of providing fair QoE in neither symmetric or asymmetric scenarios in both the local network as well as public Internet.

Based on obtained subjective scores, we conclude that the minimum hardware requirements for a three-party video conversation for the tested Web applications at the time the study was conducted were 2GB RAM and quad-core 2.5 GHz processor. The existing display size and resolution difference between smartphone configurations in this testing did not have an influence on quality ratings. This may be likely due to the fact that participants were preoccupied with always present issues such as freezing and delay, and did not focus to a great extent on image quality.

Additional functionalities

Existing WebRTC-based services differ in aspects such as the web page design, number of allowed simultaneous multiparty users, and support for additional functionalities (e.g., mute audio, hold video, recording, sharing data content, text chat). Thus, after testing each condition and providing video conversation quality ratings, participants were asked to provide their opinion with regards to additional functionalities which they considered could influence the conversation and enhance the service and overall quality. Survey results, portrayed in Table 5.13, show that the most important additional functionalities which (all) participants reported

Table 5.13: Number of participants indicating the need to include additional collaboration functionalities in WebRTC applications.

Additional functionalities	No	Yes	Percentage %
Link sharing via other service (e.g., Facebook, Email)	2	28	93.33
Mute audio	0	30	100
Hold video	0	30	100
Recording	3	27	90
Sharing data content	3	27	90
Text chat	4	26	86.67
Zooming participant	14	16	53.33
Moving participant on the screen	19	11	36.67
Name listed under the participant	13	17	56.67
Screen snapshot	6	24	80
Room lock	2	28	93.33

should be included in a WebRTC application are muting an audio stream, and disabling a video even in a private context.

All 30 participants said that the link name for entering a virtual meeting room must be easy to remember. Although a unique room link prevents uninvited guests to enter a video call, it represents a great hassle if participants do not use another service to share it. Hence, the room lock option was reported as desirable by 93.33% of participants. 90% of participants considered video recording as a useful functionality. The same percentage of participants would like to share data content. To overcome audio and video impairments and retain communication, text chat as an enhancement functionality would be appreciated by 86.67% of users. Screen snapshot is following with 80%.

While our initial study (US1) involved asymmetric end user device conditions, we later opted to avoid test design complexity caused by the influence of different devices. Consequently, after the first study, we started to use a symmetric setup, so as to maintain a similar quality of captured and reproduced audio and video at each participant. In all subsequent studies, we therefore preset the same quality per outgoing streams for all participants. To further decrease the potential impact of contextual factors, participants were further not able to select or customize the layout of the application.

Summary of key findings

The reported study US1 has shown that perceived quality can be improved on three levels:

- Firstly, mobile devices need high processing capabilities to meet the high CPU requirements imposed by audiovisual telemeetings.
- Second, the processing burden may be pushed to a centralized conferencing server in order to free up client resources.
- Third, participants would appreciate additional functionalities such as the possibilities for recording a conversation, sharing data content, or texting.

5.2 User study US2 - Impact of video encoding parameters: bitrate and resolution

Assuming a three-way video call where each participant sees three video streams on their smartphone device (the other two participants and their own video stream), user study US2 (reported in [40]) aimed to answer the following question: *what video resolutions are needed to achieve satisfactory QoE under different bitrate constraints?* This question helps to answer the high-level research question RQ2 as defined in Figure 1.1.

The study was conducted in a controlled lab environment, with all streams transmitted via the Licode[5] media server connected via a local network, and using (at that time modern) 3 GB smartphones (Figure 5.7). Licode is a platform based on WebRTC technology and enables a user to create, initialize, and publish a stream when connected to a room. The Licode architecture is based on two main components, a client API *Erizo*, responsible for signaling and handling connections to virtual meeting rooms and streams in web applications, and a video conference management API *Nuve* responsible for room management and user access control. MongoDB is used by Nuve to store information about rooms and tokens, while Erizo Controler manages and controls signaling and data streams for the rooms assigned to it by Nuve. New started ErizoControllers are automatically discovered by Nuve as long as they are connected to the same RabbitMQ instance. Distributed MCU inlcudes Erizo Agent and ErizoJS, whereby Erizo Agent is in charge of starting new processes and ErizoJS is a single broadcaster. RabbitMQ is a message broker which enables the distribution of the architecture and handles all the messages among the components of Licode, but does not handle media or communicate with the clients. With Licode media server we were able to set video parameters (resolution and bitrate).

[5] https://lynckia.com/licode/

5.2.1 Methodology

Experiments included subjective end user assessments with the goal being to investigate the impact of different video resolutions and bitrate constraints on QoE. Participants used the same WebRTC app and had the same smartphone configuration. The rational for using symmetric conditions was to eliminate the impact of different device and network settings between participants. The three-party video call was set up using a WebRTC application running on the Licode server installed in a local network, to avoid impairments caused by a commercial network, while still enabling us to control application configuration parameters, video bitrate, and video resolution (Figure 5.7).

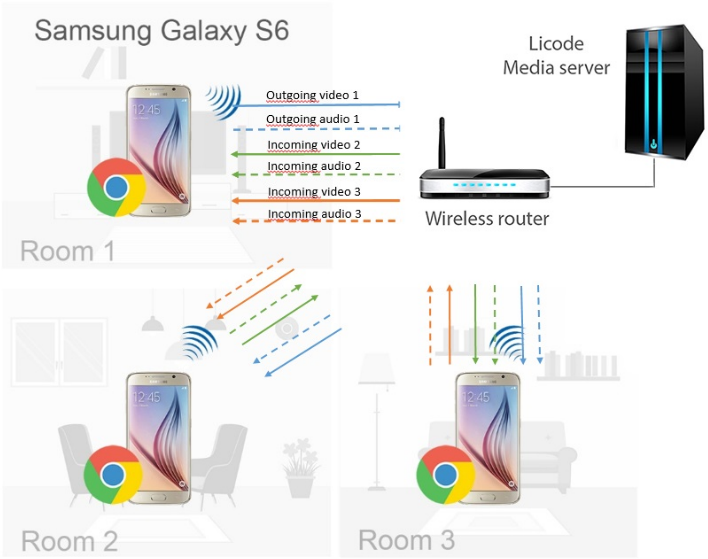


Figure 5.7: System set-up over LAN with Licode media server and Samsung Galaxy S6 (user study US2).

For our testbed setup, Licode was installed on a laptop with Intel Core i5 Processor, 2.6 GHz, 8 GB RAM and Ubuntu 12.04 LTS. The LAN connection between end user devices and the media server is Wi-Fi 802.11n, on port 3001. Video conversation was initiated through the Samsung browser version 4.0.10-53.

To explore the effects of video resolutions and bitrate limitations on perceived quality, and to avoid the impact of end user devices, all participants used the same high end smartphone configuration. Overall 108 tests were performed. The test schedule consisted of each user group (consisting of three participants) testing 12 conditions with different combinations of video resolutions and bitrates, each lasting 3 minutes. Video bitrate and video resolution in the tests were controlled using settings in Licode. We performed tests in which three video resolution were altered: 640x960 px, 480x640 px, 320x480 px under bitrate constraints so as to evaluate QoE differences under each resolution. Constant encoding bitrate constraints (assigned per resolution) were as follows: 300 kbps, 600 kbps, 1200 kbps, and 50000 kbps (corresponding

Table 5.14: Highest measured values of packet loss and jitter per test condition.

Bitrate	Resolution [px]	Packet loss [%]	Max jitter [ms]	Mean jitter [ms]
300 kbps	320x480	0.1	37.47	7.7
	480x640	0.02	27.27	7.75
	640x960	0.01	28.06	12.18
600 kbps	320x480	0.3	43.63	8.64
	480x640	0.1	34.34	9.16
	640x960	0.2	40.55	13.06
1200 kbps	320x480	0.3	41.23	13.65
	480x640	0.02	43.31	16.53
	640x960	0.32	40.92	15.45
50000 kbps	320x480	0.2	55.94	10.45
	480x640	0.6	55.94	12.82
	640x960	0.15	38.69	15.6

to “unlimited” bitrate). The process of setting up and session teardown was carried out by the administrator. After the completion of each condition, subjects were asked to rate *overall quality* and *interactivity* using a paper questionnaire and the 5-pt. ACR scale.

The physical parameters during testing were slightly different across participants (in terms of background, background light intensity, background noise level and room dimension), since each participant was located in a separate room. The maximum recorded RTT time from the media server to all client devices was 55.94 ms. We further noted packet loss and jitter from analysis of the RTP stream measured with Wireshark[6]. The highest measured values for each condition are shown in Table 5.14.

Actual throughput values (measured with Wireshark) per test condition are shown in Table 5.15. Video streams with preset bitrates of 300 kbps, 600 kbps, and 1200 kbps utilized on average approximate throughput rates of 400 kbps, 650 kbps, and 1000 kbps, respectively, to traverse the link. Throughput values measured in all test cases where the encoding bitrate was set to 50000 kbps (i.e., unlimited), were significantly lower than this maximum value, with the average throughput value measured to be approximately 1200 kbps. This means that video streams were encoded with less than 1200 kbps, despite the preset target bitrate value of 50000 kbps. At the time that this study was conducted, we did not have in-depth insights into how video quality adapted during the session (such information was analyzed in subsequent studies by accessing performance statistics available via the *webrtc-internals* tool).

Twenty-seven participants took part in the study and were divided into 9 fixed groups with

[6] https://www.wireshark.org/

Table 5.15: Average measured throughput values per test condition (test conditions differing in preset resolution and encoding bitrate). Throughput values are given in kbps.

Resolution [px] / Encoding bitrate [kbps]	300	600	1200	50000
320x480	417.03	692.40	1096.46	1168.49
480x640	405.02	644.34	1009.51	1261.10
640x960	411.39	615.23	1073.74	1213.57

3 members each. 14 male and 13 female participants took part in the studies, with an average age of 38 years (minimum 32 and maximum 65 years old). Considering acquaintances between users, free conversation was chosen to represent a natural interactive conversation. The conversations were all conducted in the Croatian language, as this was the native language to all participants.

The selected participants had no special knowledge of AV technology nor were they technical experts regarding the equipment and services to be tested. However, eight of them had participated previously in subjective assessment studies. Participants were comprised of volunteers, acquaintances, all having normal or corrected vision and normal hearing.

5.2.2 Results

Results showed that the highest streamed video resolution and video bitrate yielded the lowest MOS scores for all test cases. Figure 5.8 depicts the dependency of overall quality ratings on different combinations of values for bitrate and resolution parameters. Two main conclusions can be drawn:

1. the resolution 640x960 px should not be set for any of the tested bitrate limitations, as for that resolution MOS scores for overall quality are always below 4 and lower than other tested resolution settings;
2. tagret bitrate setting 50000 kbps results in significantly reduced user perceived overall quality for all resolutions above 320x480 px, meaning that the capabilities of the tested mobile phones had trouble processing multiple real-time videos with high bitrates and resolutions.

These results are in line with our previous findings and may be considered generally applicable, as for the testing procedure powerful Samsung Galaxy S6 mobile phones were used, which were considered high end mobile devices available on the market at the time of the study. In every test with preset video encoding bitrate to 50000 kbps, participants reported picture freezing, although the speech was acceptable, so communication was not completely interrupted. Consequently, the ratings were still fair, although significantly lower as compared to the other bitrate limitations. The area with an optimal combination of parameters is clearly depicted in Figure

5.8 with MOS scores over 4.5. What is interesting is that all combinations between 1200 and 300 kbps and both 320x480 and 480x640 px resolutions are in this area. In experiments with video bitrate limitation of 300 kbps, overall quality gained the highest scores for all resolutions. The experiments with a resolution preset to 480x640 px gained the highest average scores (over 4.5) for overall quality.

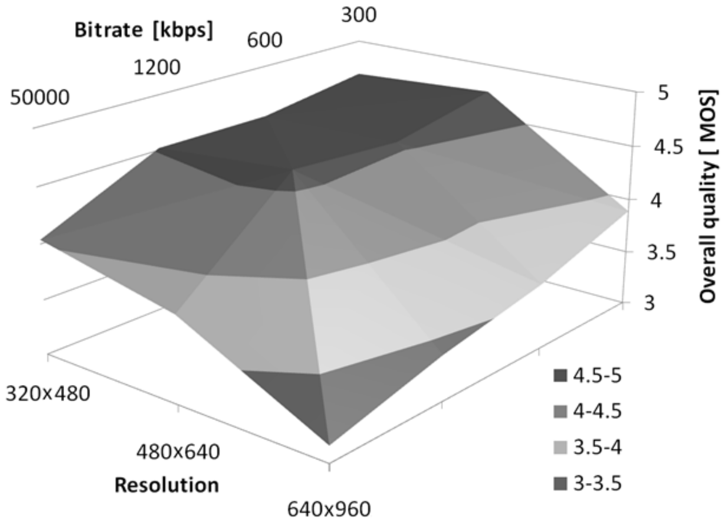


Figure 5.8: Overall quality for each combination of encoding bitrate and resolution settings.

Besides overall quality, we also measured interactivity perceived by the users. In Figure 5.9 we depict the MOS values for both overall quality and interactivity for resolution 480x640 px across all bitrate limitations. It can be noted that overall quality and interactivity are highly correlated and that their 95% confidence factors overlap for every experiment.

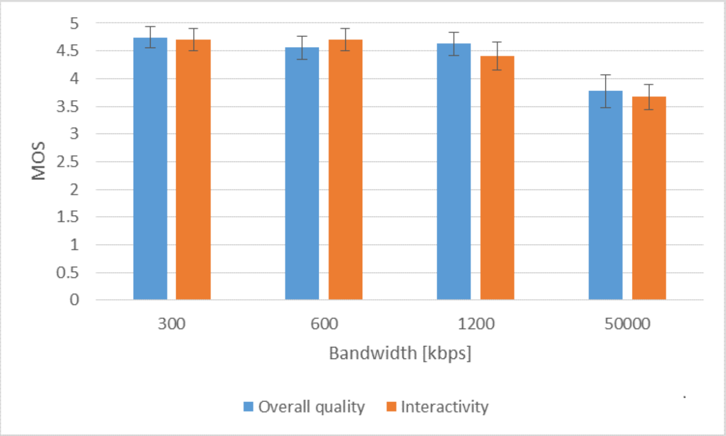


Figure 5.9: Overall quality and interactivity for 480x640 px resolution across all bitrate values.

Instead of achieving the highest quality ratings, the largest test resolution as well as highest bitrate seem to have caused congestion on the smartphones, which ultimately affected the perceived quality. The amount of generated traffic had a significant influence on the QoE, especially for bitrates higher then 1200 kbps for each resolution tested, which may be attributed to high demands on smartphone processing power. However, despite the lack of smartphone processing power, 5.1” screen size with the corresponding display resolution remains as an argument that resolutions higher than 480x640 px are unnecessary for three-party video video calls.

Summary of key findings

Based on the presented results in this section, the following key findings can be highlighted for user study US2:

- Streaming at a resolution of 640x960 px or higher in the context of video calls on smartphone devices may be considered unnecessary. Even a resolution of 480x640 px will often be reduced by the application due to CPU overuse, and as such may be considered unnecessary for smartphones. With respect to bitrate limitation (target output video bitrate), 600 kbps is also a rate which may be too high and will depend on the mobile device configurations. Subsequent studies are needed to further investigate acceptable bitrate values in the context of a three-party telemeeting on mobile devices.

5.3 User study US5 - Establishing a lower threshold for setting acceptable video resolutions and bitrates

The goal of user study US5 (reported in [42]) was twofold: first, to answer what video resolutions and bitrates are needed to achieve acceptable QoE, and accordingly to identify lower thresholds of encoding parameters (described in this section); and secondly, to analyze objective video quality metrics calculated using screen recordings of multiparty video calls, namely blurriness and blockiness, and their correlation with perceived video quality reported by call participants (described in the following section 5.4.1). This user study helps to answer the high-level research questions RQ2, RQ3 and RQ4 as defined in Figure 1.1.

5.3.1 Methodology

Measurements involving interactive three-party audiovisual conversations carried out in a leisure context were conducted in a controlled laboratory environment (one participant per site) over a Wi-Fi network, and with symmetric device conditions. In the experiments, video resolution, bitrate, and frame rate were predefined using settings on the Licode server. Licode was installed in a local network on a computer with Intel Core i5 Processor, 2.6 GHz, 8 GB RAM and Ubuntu 14.04 LTS (Figure 5.10). Participants took part in the call using Samsung Galaxy S7 smartphones with 4 GB of RAM. During each call, the smartphone screen was recorded using the DU recorder application[7] at 1088x1920 px, and about 22 fps, as MPEG-4 (Base Media / Version 2), AVC. To monitor video quality and service performance, WebRTC session-related data was collected via Chrome browser and *webrtc-internals* [147]. *Webrtc-internals* is an internal functionality for collecting statistics (such as: round trip time, packet loss, delay, average encoding time of the frame, actual encoding bitrate, frame height, frame width, frame rate, available send bandwidth) about ongoing WebRTC sessions. To obtain statistics, a session has to be opened in the Chrome browser, and while in that session, another tab has to be open with the following URL: *chrome://webrtc-internals*. Before terminating the session, a dump file can be generated and downloaded.

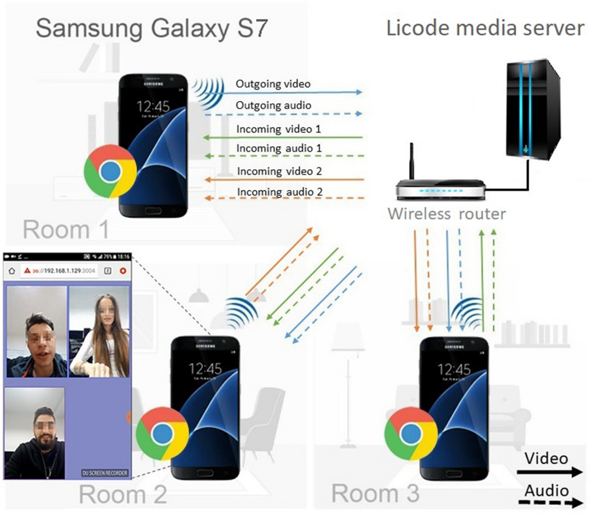


Figure 5.10: Testbed set-up over a LAN connection (user study US5).

The test schedule consisted of 7 testing conditions, with videos encoded with the VP8 video codec, and resolutions, bitrate, and frame rate set according to Table 5.16. Each test condition was evaluated by 9 groups, leading to a total of 63 performed tests[8]).

The setup was symmetrical for all participants within each group. Established video telemeetings lasted for two minutes per test session and were initiated through a WebRTC application within the Google Chrome 63.0.3239.111 browser.

Twenty-seven participants (20 female and 7 male) took part in the study on a voluntary basis, with an average age of 22 years (min age 20, max. age 23). Participants were divided into nine groups, formed based on acquaintances. All participants were students, non-experts in the AV field, and had previous experience with applications such as Skype, Viber, and WhatsApp.

[7] https://du-recorder.en.uptodown.com/android

[8] We discarded the data from one group due to erroneous measurements or incomplete responses.

Table 5.16: Test conditions used in user study US5.

Test conditions	Video resolution [px]	Frame rate [fps]	Bitrate [kbps]
Test case 1 (TC1)	180x240	15	200
Test case 2 (TC2)	360x480	15	300
Test case 3 (TC3)	240x360	15	150
Test case 4 (TC4)	120x180	15	100
Test case 5 (TC5)	240x360	15	200
Test case 6 (TC6)	120x180	20	200
Test case 7 (TC7)	240x360	20	300

5.3.2 Results

WebRTC internals data and MOS values: To check the actual *sent* and *received* video qualities, and to be sure that participants were in fact rating the preset quality levels (as opposed to some dynamically adapted levels) we analyzed *webRTC-internals* data. We observed that resolution adaptation occurred only in TC2 within 6 video streams due to CPU overuse (Table 5.17). We note that this adaptation is automatically invoked by the application. In those cases, resolution was decreased to 270x360 px, and lasted at this level for an average of 50.45% of the session time. Within all test cases, packet loss was very low (around 0.001%). Only in TC7, within one group, packet loss yielded 0.96%.

If we want to avoid CPU overuse which participants can detect, we conclude that video settings used in TC2 may be preset as an upper bound in terms of resolution, frame rate, and bitrate, when used in the context of three-party video calls established using smartphones with processing capabilities comparable to those tested (4GB of RAM). On the other hand, while participants provided the highest average quality ratings for TC2, we see that only a slight decrease in average ratings is observed in the case of TC5, albeit TC5 involved resolution set to 240x360 px, the same frame rate, and 200 kbps bitrate (rather than 300 kbps as used in TC2). It is thus worth considering whether the significant increase in resources (from TC5 to TC2) is worth the only slight gain in perceived quality.

We further conclude that the test case with the lowest video quality (TC4: 120x180 px, 15 fps, 100 kbps) is not a recommendable settings for a three-party video calls, with subjective ratings giving an average of 3.17 for audio quality, 2.33 for video quality, and 2.83 for both synchronization and overall quality. We observed that the cause of such low ratings is not actually the resolution, but rather insufficient bitrate. TC6, which had the same resolution, but a slightly higher frame rate (20 fps) and higher available bitrate (200 kbps), resulted with a video MOS of 3.13 and overall MOS 3.38.

Table 5.17: MOS ratings and WebRTC internals statistics of mean values per test condition.

Test case	TC1	TC2	TC3	TC4	TC5	TC6	TC7
Percentage of session time where actual streamed resolution corresponded to the set resolution	100%	86.28 %	100%	100%	100%	100%	100%
Percentage of session time where actual streamed frame rate corresponded to the set frame rate and +/-1	96.22%	93.84 %	94.99%	97.06%	96.72%	91.67%	88.12%
AVG frame rate	14.91	14.82	14.85	14.91	14.92	19.78	19.53
MOS Audio quality	3.67	3.83	3.21	3.17	3.67	3.46	3.50
MOS Video quality	3.50	3.75	3.17	2.33	3.67	3.13	3.58
MOS AV synchronization	3.46	3.63	3.17	2.83	3.63	3.29	3.50
MOS Overall quality	3.63	3.75	3.17	2.83	3.63	3.38	3.50

Summary of key findings

Based on the results of study US5, we summarize the following findings:

- Occasional video impairments did not significantly impact overall perceived quality.
- Participants were not always able to distinguish and report impairments, possibly due to the small preview size, short duration and/or low strength of disturbances, or their engagement in the conversation.

5.4 User studies US5 and US6 - investigation of the relationship between objective quality metrics and subjective quality ratings

5.4.1 User study US5

Digital video systems can add edges (e.g., blocking) or reduce edges (e.g., blurring). Blocking distortion can be introduced by coding and/or transmission errors (when the video encoder is not able to process the whole stream) [148]. Perceived video blurriness appears when a loss of spatial details or sharpness at edges or texture regions in the image occurs [101]. Blurriness can appear during fast camera movement or when capturing high movement content. Wrong focus, inadequate resolution, and issues with video compression are also factors that can contribute to the video blurring. Video compression methods are based on the frequency transformation followed by a quantization process that often discards coefficients with low amplitudes. While the energy of natural visual signals is concentrated at low frequencies, quantization reduces high frequency energy which will result with the blurriness occurrence in the reconstructed signal. On the other hand, a slow Internet connection speed and limited bandwidth can cause video quality degradation (in terms of resolution and bitrate) and consequently blurriness. Hence, in the case of video coding, a fundamental trade-off happens between image quality (distortion), compression (rate), and computational complexity. To analyze objective video quality, we used the MSU Video Quality Measurement Tool (VQMT) Professional Version 10.2.[9] (a screenshot of measurements recorded using the tool is given in Figure 5.11).

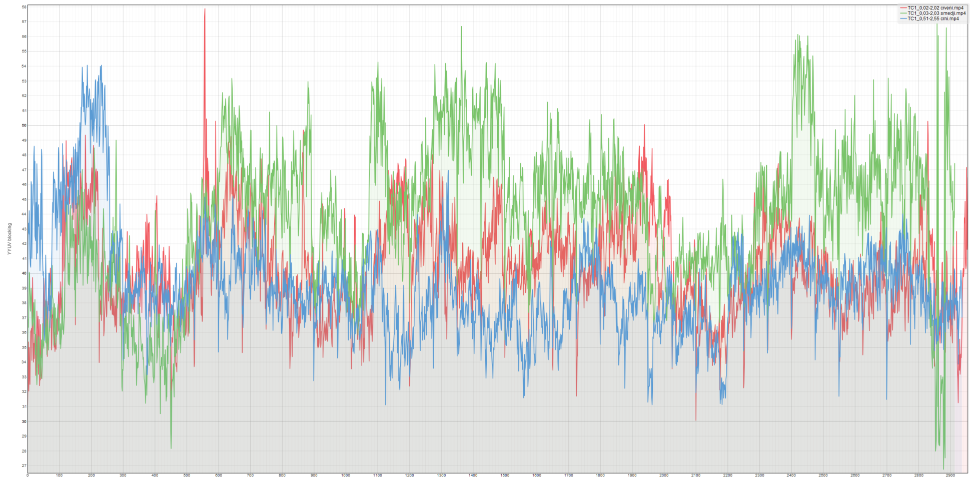


Figure 5.11: Example of blockiness measurements collected using the MSU video quality measurement tool during a three-party video call. The horizontal axis indicates frame number, while corresponding per-frame blockiness values (shown per participant: red-, green-, blue lines) are plotted on the vertical axis.

To be able to compare objective measurements with subjectively perceived impairments, participants were asked to report whether or not they experienced blurriness and blockiness, and whether or not they noticed any video freezes (this data was collected following each test case). While 66.67% participants responded that they noticed blurriness in the test case with the lowest video quality and lowest ratings (TC4), in the objectively highest video quality test case (TC2), blurriness was observed by 50% of all participants. The least number of participants reported having noticed blurriness in TC5 (240x360 px, 15 fps, 200 kbps) with a share of 45.83% (Table 5.18). Participants reported blockiness in test cases where insufficient bitrate was preset. Blockiness was reported in TC1 only by 8.33% participants, while in TC4 and TC7 by 37.5%.

[9] https://www.compression.ru/

Table 5.18: Percentage of participants reporting disturbances.

Test case	Blurriness	Blockiness	Freezes
TC1	62.50%	8.33%	4.17%
TC2	50.00%	4.17%	0.00%
TC3	50.00%	29.17%	45.83%
TC4	66.67%	37.50%	25.00%
TC5	45.83%	16.67%	8.33%
TC6	62.50%	25.00%	8.33%
TC7	41.67%	37.50%	29.17%

Table 5.19: Mean values of video impairments and rated video quality (VQ).

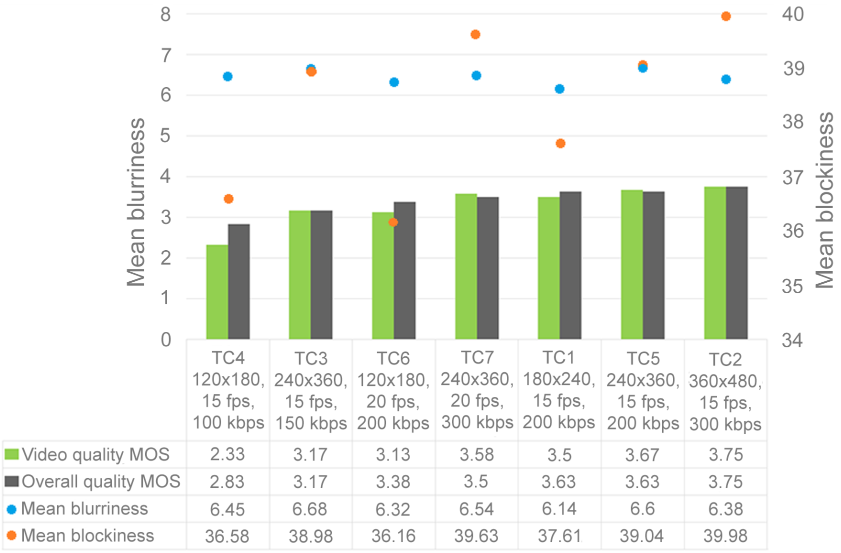
Test case	Blurriness median/mean/StDev	Blockiness median/mean/StDev	MOS VQ	%GoB VQ
TC1	6.10 / 6.14 / 0.34	36.90 / 37.61 / 5.14	3.5	50%
TC2	6.29 / 6.38 / 0.45	39.41 / 39.98 / 4.84	3.75	66.67%
TC3	6.72 / 6.68 / 0.61	38.35 / 38.98 / 6.21	3.17	37.5%
TC4	6.40 / 6.45 / 0.53	36.27 / 36.58 / 4.53	2.33	4.17%
TC5	6.61 / 6.60 / 0.68	38.44 / 39.04 / 5.18	3.67	62.5%
TC6	6.29 / 6.32 / 0.43	35.75 / 36.16 / 3.83	3.13	33.33%
TC7	6.52 / 6.54 / 0.88	38.48 / 39.63 / 7.9	3.58	66.67%

Based on our results, it turns out that short video freezes did not have a significant impact on reported perceived quality. In fact, only in six sessions (out of a total of 58 sessions), two participants reported having noticed a video freeze. In all other sessions where video was reported as being frozen, this was noticed by only one participant from the session. TC2 is the only scenario where participants did not report any freezes. In the other cases, 4.17-29.17% of participants reported freezes.

Blurriness and blockiness per test case

Based on results obtained during the video calls, we wanted to investigate the relationship between objective no-reference video metrics, namely blurriness and blockiness, and subjective user ratings. We note that higher measured values of blockiness and blurriness indicate better video quality. A summary of results is given in Table 5.19 and Figure 5.12.

Mean opinion score is considered to be one of the most straightforward evaluation measures for subjective quality assessment, and it would be sufficient if the rating distribution is normal (bell curve). However, results can show skewness, thus additional metrics can provide missing important information with respect to rating distributions [124]. For service providers, insights into the ratings distribution and the usage of the additional metrics besides mean, such as "Good or better" (%GoB) or the percentage of users abandoning a service (Terminate Early, %TME) can help with planning and management of their infrastructure. The "Good or better" ratio indicates the percentage of participants assessing the test condition as 4 "Good" or 5 "Excellent". Our calculated results show that to keep more than 60% of participants satisfied (assuming this corresponds to rating the video quality in the call with 4 or 5), scenarios where VQ MOS was above 3.67 have to be considered. This means that resolution should be at least 240x360 px, corresponding 200 kbps video bitrate and frame rate of 15 fps. A significant drop occurs when VQ MOS drops below 3.17, at which level only 37% of the participants were satisfied.



	TC4 120x180, 15 fps, 100 kbps	TC3 240x360, 15 fps, 150 kbps	TC6 120x180, 20 fps, 200 kbps	TC7 240x360, 20 fps, 300 kbps	TC1 180x240, 15 fps, 200 kbps	TC5 240x360, 15 fps, 200 kbps	TC2 360x480, 15 fps, 300 kbps
🟩 Video quality MOS	2.33	3.17	3.13	3.58	3.5	3.67	3.75
■ Overall quality MOS	2.83	3.17	3.38	3.5	3.63	3.63	3.75
● Mean blurriness	6.45	6.68	6.32	6.54	6.14	6.6	6.38
● Mean blockiness	36.58	38.98	36.16	39.63	37.61	39.04	39.98

Figure 5.12: Mean values of blurriness and blockiness with associated video and overall quality MOS scores per each test condition.

If we compare TC1 (180x240 px, 15 fps, 200 kbps) and TC7 (240x360 px, 20 fps, 300 kbps), we observe that MOS was higher in TC1 than TC7, for all rated quality dimensions except for video quality (which was only slightly lower). In terms of determining video codec configuration parameters, it may thus be possible to save 100 kbps, avoid possible CPU overuse, and still obtain a higher average score for overall quality. TC1 was preset with a lower resolution then TC7, which participants noticed, but did not have a significant impact when rating other aspects.

Distributions of blurriness and blockiness values for different subjective video quality ratings

With the summary statistics, a wide range of values overlapped across different user ratings. Thus, to gain better insights and to visualize data and performance indicators, we used histograms to measure how frequently values appear in our data sets. The histograms for user video quality (VQ) ratings of 1 and 5 have a notably different spread and correlated frequency of values compared to VQ 3 or 4, since video quality was rated as "Bad" in only 2.64% cases, and as "Excellent" in only 7.93% cases.

The following histograms show the blockiness and blurriness values from all test scenarios associated with corresponding video quality ratings (Figure 5.13 and Figure 5.14). We split data into 20 bins for blockiness values and 10 bins for blurriness values. We chose a different number of bins in order to show underlying patterns and data trend. Each bin contains the frequency of occurrences of values in the data set that are contained within that bin. On the graphs, we can observe shifted distributions to the right per higher VQ rating for both blockiness (Figure 5.13) and blurriness (Figure 5.14), which correlates to better quality.

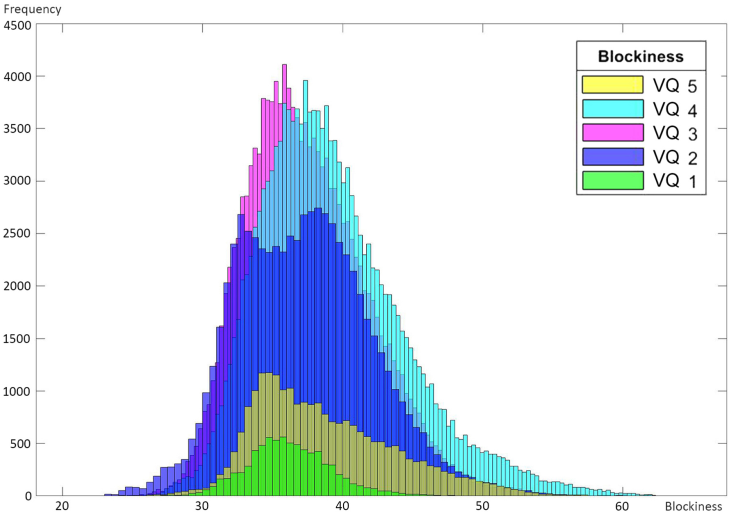


Figure 5.13: Frequency of blockiness values per frame for video quality (VQ) user ratings. Note: VQ was rated per video call, so all frames belonging to a given call are colored the same, i.e., according to the VQ rating.

Comparing blockiness and blurriness graphs, blurriness values are more inconsistent and spread due to the camera movement and participants moving around, which impacted the blurriness. Thus, to better describe sample data we fitted blockiness and blurriness values to common distributions using MATLAB R2018b[10].

We evaluated (based on log likelihood values and probability plots) that the best fit for blockiness is the Birnbaum-Saunders distribution (Table 5.20). Birnbaum-Saunders distribution is defined with the beta (scale) parameter and gamma (shape) parameter. Since video quality was most often rated with "Fair" or "Good", for those two ratings we have the largest value set, and consequently the largest value span. Therefore, fited distributions with respective probability plots have the longest tales (Figure 5.15).

Mean values of fitted data samples are in ascending order from video quality user rating VQ 1 to VQ 3, while VQ 5 value is placed between VQ 3 and VQ 4. One of the possible reasons

[10] https://www.mathworks.com/

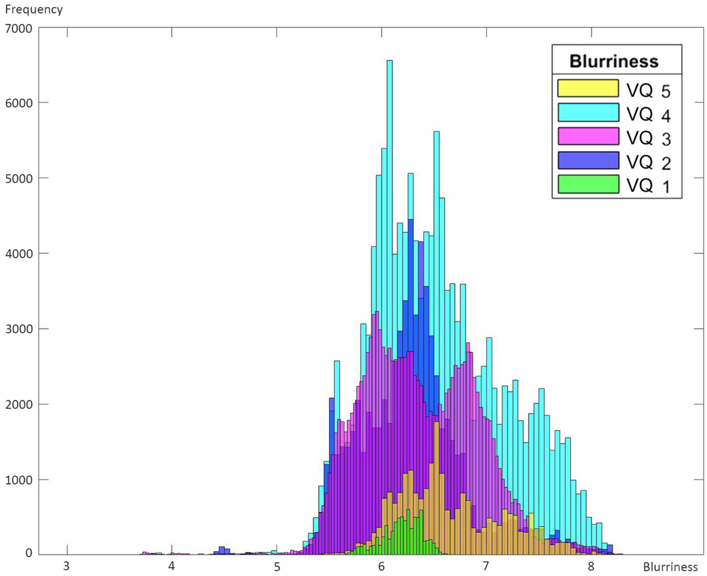


Figure 5.14: Frequency of blurriness values per frame for video quality (VQ) user ratings. Note: VQ was rated per video call, so all frames belonging to a given call are colored the same, i.e., according to the VQ rating.

Table 5.20: Measured blockiness values per frame per video quality user ratings with fitted Birnbaum-Saunders distribution.

Blockiness per VQ user rating	1	2	3	4	5
Mean	36.1858	37.282	37.488	39.254	38.744
Variance	8.910	23.663	18.8266	26.151	24.651
Parameter betaestimate	36.063	36.968	37.240	38.925	38.430
Parameter betaStd. Err.	0.03389	0.01840	0.01165	0.01329	0.03202
Parameter gamma estimate	0.08242	0.13021	0.11555	0.13001	0.12789
Parameter gamma Std. Err.	0.00066	0.00035	0.00022	0.00024	0.00059

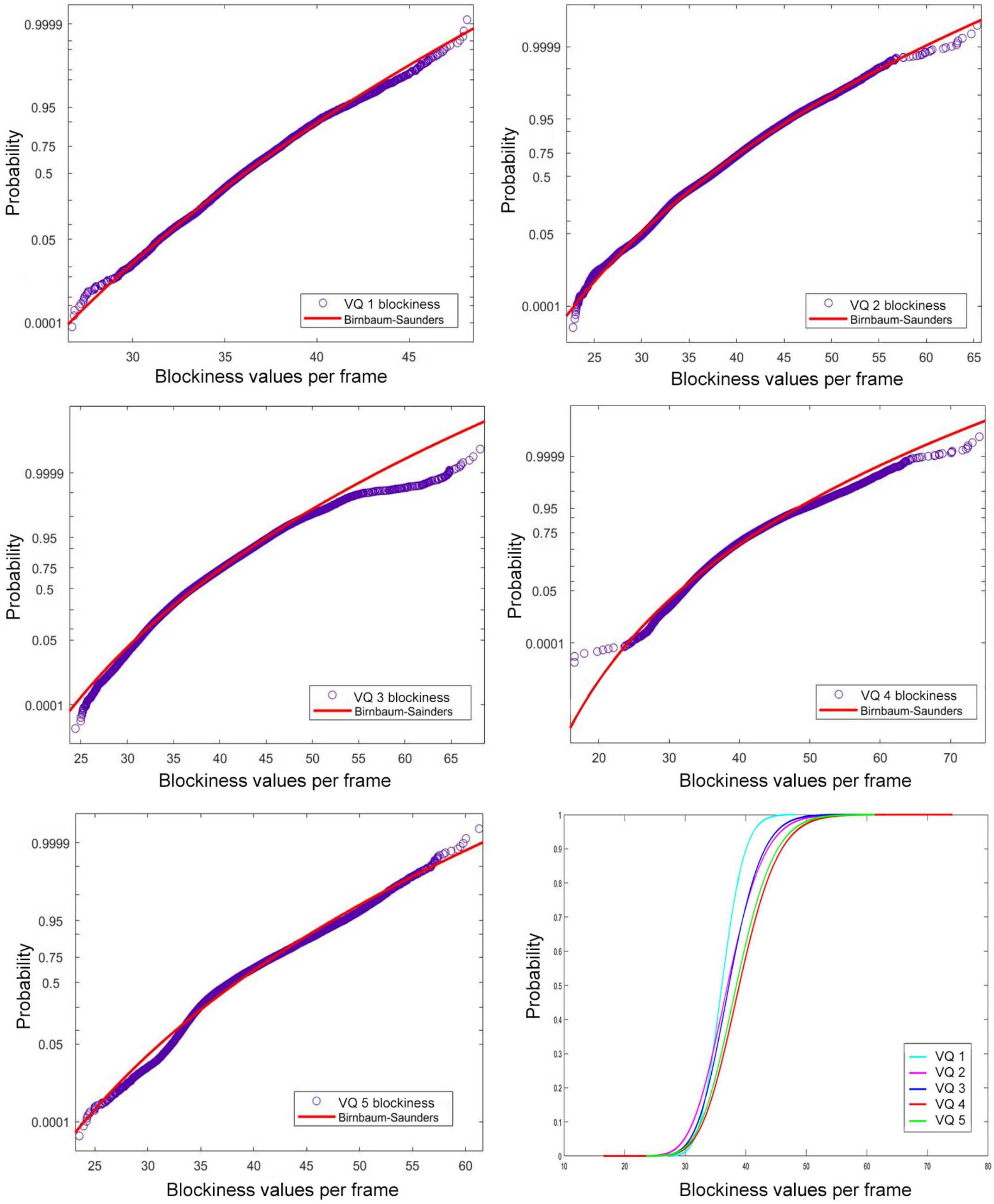


Figure 5.15: Probability plots for Birnbaum-Saunders distribution for blockiness values per frame per video quality user rating.

Table 5.21: Distribution of blurriness values per frame and per video quality rating level, with fitted Burr distribution for VQ 1 and VQ 2 user ratings, and Gamma distribution for VQ 3, VQ 4, and VQ 5 user ratings.

Blurriness perVQ user rating	**1**	**2**	**3**	**4**	**5**
Mean	6.18757	6.28235	6.36187	6.53864	6.6438
Variance	0.03831	0.2781	0.29597	0.38999	0.24336
Parameter alpha estimate	6.6488	6.15611	-	-	-
Parameter alpha Std. Err.	0.04859	0.00485	-	-	-
Parameter c estimate	40.9839	24.1488	-	-	-
Parameter c Std. Err.	0.62142	0.13621	-	-	-
Parameter k estimate	11.4123	0.79168	-	-	-
Parameter k Std. Err.	2.88852	0.00941	-	-	-
Parameter a estimate	-	-	136.744	109.626	181.378
Parameter a Std. Err.	-	-	0.52379	0.40842	1.67562
Parameter b estimate	-	-	0.04652	0.05964	0.03662
Parameter b Std. Err.	-	-	0.00017	0.00022	0.00033

could be due to the significantly smaller number of sample data inputs. The highest yielded blockiness measured value for VQ 1 was 48.15, VQ 2: 65.35, VQ 3: 68.15, VQ 4: 74.11, while for VQ 5 it was 61.28. While we can observe a positive trend, we do not observe high consistency, partly because of a large difference between sample set sizes. This trend could also be due to the peak level of annoyance experienced by users at a certain blockiness level, which could later settle to a slightly better QoE beyond this blockiness level owing to saturation effects related to user QoE.

For the blurriness values (measured per frame) values from sessions where video quality was rated with “Bad” and “Poor” were fitted to a Burr distribution, while values corresponding to sessions rated as “Fair”, “Good”, or “Excellent” were fitted to a Gamma distribution. A Burr distribution is defined with three parameters: alpha-scale parameter, c-first shape parameter, and k-second shape parameter. Gamma distributions is defined with a-shape parameter and b-scale parameter.

Figure 5.16 shows probability plots for blurriness values rated with VQ 1 to 5, collected during sessions across seven different test cases. Results are summarized in Table 5.21. Mean values of fitted data samples for blurriness ascend in order from VQ 1: 6.18 to VQ 5: 6.64. The highest yielded blurriness measured value for VQ 1 was 6.62, VQ 2: 8.22, VQ 3: 8.46, VQ 4: 8.63, and VQ 5: 8.44.

The blurriness probability plot with fitted distributions shows some shift to the right, where

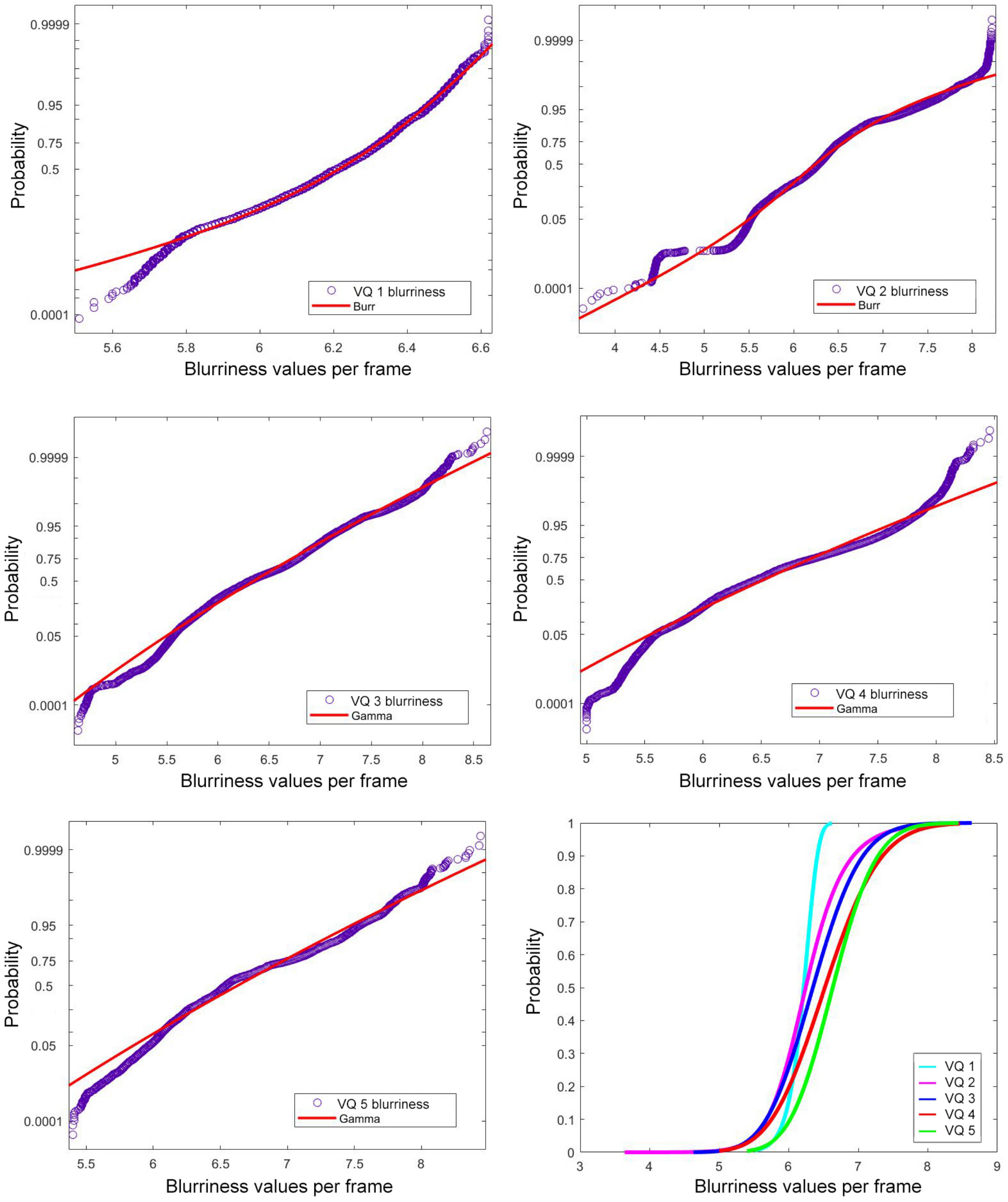


Figure 5.16: Probability plots for Burr and Gamma distributions for blurriness values per frame per video quality user rating.

better quality values correspond to higher QoE. However, to obtain more precise results, further testing should be done, with an adapted methodology to achieve more stable session performance.

Due to the similar and overlapping blurriness and blockiness values, we conclude that it is difficult to correlate specific levels of blurriness and blockiness with user ratings. However, participants did notice the changes in objective video quality and rated them accordingly (Figure 5.17). In test cases with “tighter” bitrate (enough for lower motion) for a chosen resolution, video quality scores correlated better with overall quality than audio quality scores. In test cases with assigned higher bitrates (TC1, TC6, TC7), audio quality scores correlated better with overall quality scores.

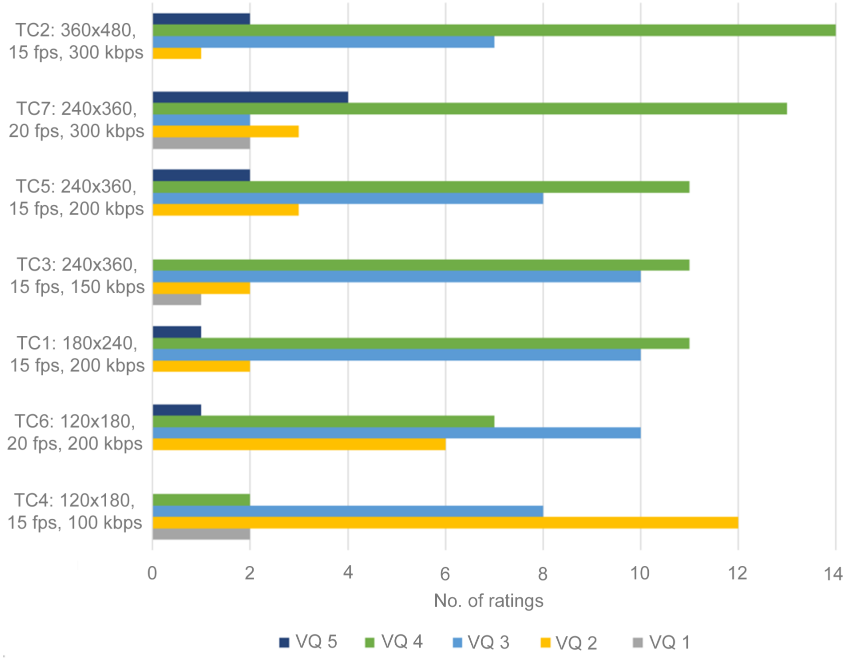


Figure 5.17: Number of occurrences of participant VQ ratings for each level of the used 5 pt. rating scale (user study US5).

Correlation between mean blurriness and blockiness values with objective metrics and subjective ratings

Finally, we used Pearson’s correlation coefficients to measure the strength of the relationship between measured objective blurriness and blockiness and perceived video quality mean ratings, resolution, and bitrate (Table 5.22). The data shows significant correlation (at the 0.01 level) only between perceived video quality and blockiness (mean and median), while a weak positive correlation was found between resolution and blurriness, and a negative correlation between bitrate and blurriness, as well as between VQ MOS and blurriness.

Taking into consideration perceived disturbances reported by participants (perceived blurriness, blockiness, and freezes), results show no significant correlation between any examined parameter (Table 5.23).

Table 5.22: Correlations between mean and median blurriness and blockiness measured objective values and resolution, bitrate and VQ MOS collected in US5.

Pearson correlation	Blurriness mean	Blurriness median	Blockiness mean	Blockiness median
Resolution	0.18	0.26	0.93**	0.89**
Bitrate	-0.17	-0.11	0.61	0.66
VQ MOS	-0.03	-0.04	0.69	0.72

** Correlation is significant at the 0.01 level

Table 5.23: Correlations between reported by participants perceived impairments (blurriness, blockiness and freeze) and resolution, bitrate and VQ MOS collected in US5.

Pearson correlation	Perceived blurriness	Perceived blockiness	Perceived freezes
Resolution	-0.68	-0.496	-0.162
Bitrate	-0.641	-0.370	-0.376
VQ MOS	-0.699	-0.633	-0.411

5.4.2 User study US6

User study US6 (reported in [43]) aimed to extend the results from the previous study US5, where we wanted to obtain clearer insights with respect to objective metrics. Thus, we further examined the impact of blurriness and blockiness on the perceived quality, along with the impact of different resolutions under constrained bandwidth. This user study helps to answer the high-level research questions RQ2, RQ3, and RQ4 as defined in Figure 1.1.

Methodology

This study involved three participants in a video call, one-per-site setup and a leisure context. Measurements were conducted in a controlled laboratory environment over a Wi-Fi network, and with symmetric device conditions. Video resolution, bitrate, and frame rate in the experiments were predefined using settings on the Licode server installed in a local network on a computer with Intel Core i5 Processor, 2.6 GHz, 8 GB RAM and Ubuntu 14.04 LTS. Participants took part in the conference using Samsung Galaxy S7 smartphones with 4 GB of RAM. During the call, the screen of the smartphone was recorded using the DU recorder application.

The test schedule consisted of 6 tests, but with four different test conditions. Videos were encoded with the resolution, bitrate, and frame rate set according to Table 5.24. We wanted to investigate the impact of different resolutions under the same bitrate constraint of 300 kbps. To

Table 5.24: Test conditions used in user study US6.

Test conditions	Video resolution [px]	Frame rate [fps]	Bitrate [kbps]
Test case 1 (TC1)	180x240	10	300
Test case 2 (TC2)	180x240	20	300
Test case 3 (TC3)	320x430	20	300
Test case 4 (TC4)	240x320	20	300
Test case 2’ (TC2’)	180x240	20	300
Test case 3’ (TC3’)	320x430	20	300

check reliability of certain test scenarios, and investigate the extent to which repeated measurements yield consistent results, we performed TC2 and TC3 twice.

Each test condition was evaluated by 9 groups (formed based on acquaintances), leading to a total of 54 performed tests. The setup was symmetrical for all participants within each group. Established calls lasted for two minutes per test session and were initiated through a WebRTC application within the Google Chrome browser. After the completion of each condition, subjects were asked to rate overall-, audio-, and video quality using a paper questionnaire and the 5-pt. Absolute Category Rating (ACR) scale.

Twenty-seven participants (11 female and 16 male) took part in the study on a voluntary basis, with an average age of 23 years (min age 22, max. age 28). The majority of participants were students (four of them were employed), non-experts in the AV field, and had previous experience with using video conference applications.

Results

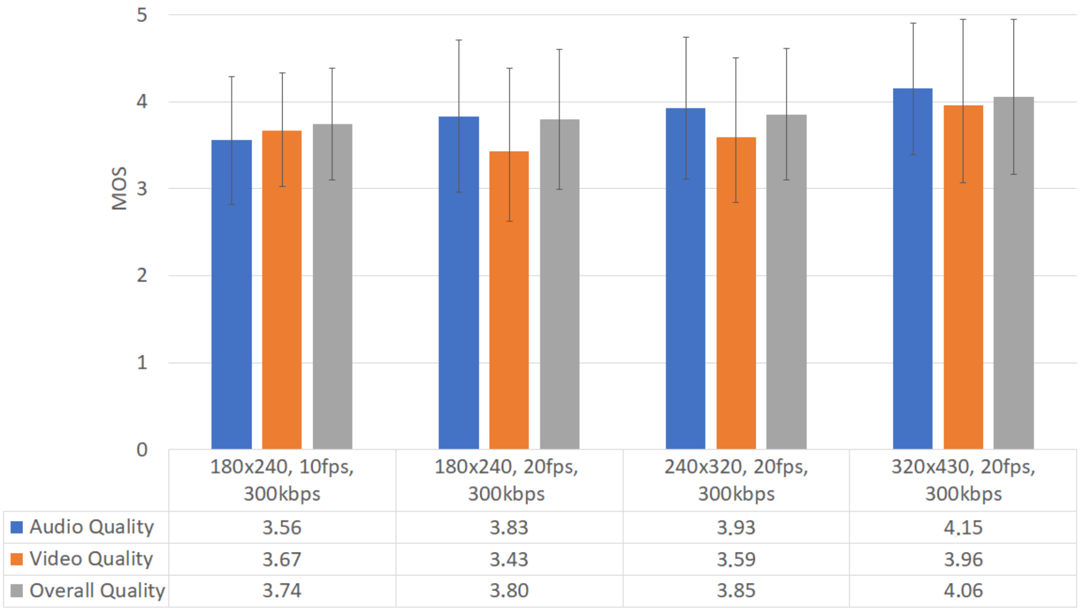
We analyzed *webRTC-internals* data to check the obtained video qualities during the sessions. Resolution adaptation occurred only in test case 320x430 px, 20 fps, 300 kbps due to both CPU overuse and bandwidth limitation, where resolution was decreased to 240x321 px (Table 5.25). Within all test cases, packet loss was almost zero, and only in a few streams packet loss yield an average of 3.07%.

Results showed that the highest rated condition per all rated qualities was 320x430 px, 20 fps, 300 kbps with an average audio quality MOS 4.15, video quality MOS 3.96, and overall quality MOS 4.06 (Figure 5.18). The lowest video quality score (3.43) yielded test condition with resolution 180x240 px, 20 fps, and 300 kbps. Comparing test condition TC1 with TC2 and TC2’ differing only by the frame rate value we can observe that overall quality was perceived as better in case where audio quality was perceived as better as well. Interestingly, test condition 180x240 px, 10 fps yielded higher mean video quality score (but lower overall) than test con-

Table 5.25: MOS ratings and WebRTC internals statistics of mean values per test condition.

Test case	TC1	TC2	TC3	TC4	TC2’	TC3’
Percentage of session time where actual streamed resolution corresponded to the set resolution	100%	100%	95.88%	100%	100%	94.21%
Percentage of session time where actual streamed frame rate corresponded to the set frame rate and +/-1	100%	86.62%	74.76%	78.61%	94.52%	86.95%
AVG frame rate [fps]	10.11	19.26	18.87	18.86	19.76	19.57
AVG bitrate [kbps]	294.51	271.31	271.52	268.56	295.07	287.58
MOS Audio quality	3.56	3.89	4.07	3.93	3.78	4.22
MOS Video quality	3.67	3.41	3.93	3.59	3.44	4
MOS Overall quality	3.74	3.81	4	3.85	3.78	4.11

dition 180x240 px, 20 fps, showing that audio component had a greater impact than perceived video quality.



	180x240, 10fps, 300kbps	180x240, 20fps, 300kbps	240x320, 20fps, 300kbps	320x430, 20fps, 300kbps
■ Audio Quality	3.56	3.83	3.93	4.15
■ Video Quality	3.67	3.43	3.59	3.96
■ Overall Quality	3.74	3.80	3.85	4.06

Figure 5.18: Aggregated audio-, video-, and overall quality MOS results for each unique combination of encoding settings.

The median for video quality across all test cases was 4. In test cases TC1, TC2, and TC4, 75% of reported ratings were 3 and 4, while the value for the quartile 1 (Q1) was 2.5 for TC3 and TC2’, and 4 for TC3’ (Figure 5.19). Only TC3’ had a third quartile value 5, while the rest of the test cases had 4. Mild outliers were found towards only to the bottom of the box plot. Even though the distribution of the VQ ratings was different between TC3 and TC3’, mean VQ ratings did not differ significantly 3.93 (TC3) comparing to 4 (TC3’). We can attribute that to the perceived impairments of blockiness, blurriness, and freezes which were reported more often in TC3.

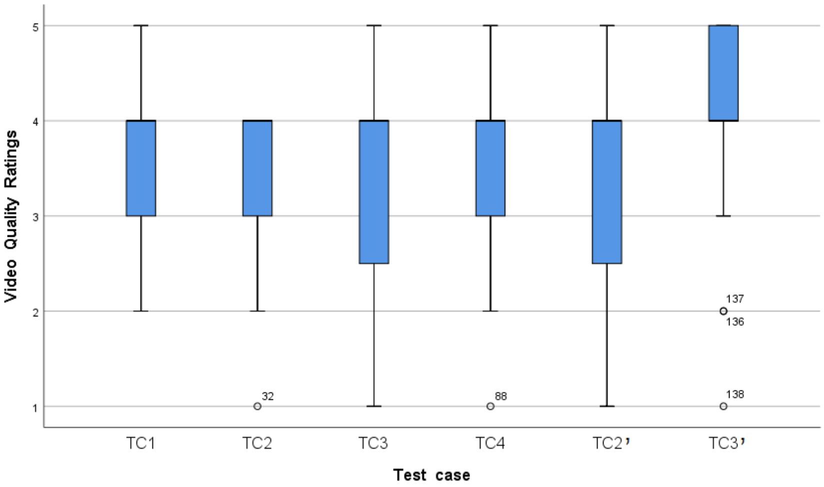


Figure 5.19: Box plot of video quality ratings per each test case.

We can conclude that 300 kbps bitrate should be enough to transmit resolutions up to 320x430 px without impairment caused by CPU overuse or bitrate limitation, in the context of three-party video calls established using smartphones with processing capabilities comparable to those tested (4GB of RAM). We further conclude that a resolution of 180x240 px should still provide at least fair QoE, hence it may be set in case of limited resources.

Since with our previous study US5 we did not obtain clear results with respect to whether or not blurriness or blockiness are useful objective metrics (in terms of estimating subjective ratings) in case of a multiparty video call on mobile devices, we wanted to analyze the data one more time within this study. Thus, we asked participants to report whether or not they noticed blurriness, blockiness, or any video freezes. Participants reported that they noticed blurriness in the test case with the lowest video quality (TC2) by 55.55%, while in the objectively highest video quality test case (TC3') the least number of participants (29.63%) reported having noticed blurriness (Table 5.26). On the other hand, in test cases with the same resolution 180x240 px TC1 and TC2', only 14.81% participants reported blockiness, while in Tc3 the highest number of participants (25.93%) reported blockiness. Similar to the findings of previous studies, the current study results showed that short video freezes did not have a significant impact on reported perceived quality. Participants reported sessions as being frozen in all test cases at least once (3.71% to 14.81%) or several times (3.71% to 11.11%).

Based on results obtained during the video calls, we wanted to further investigate the relationship between blurriness and blockiness, and subjective user ratings, whereby better objective video quality is achieved by higher measured values of blockiness and blurriness. A summary of results is given in Table 5.27.

Additionally, calculated %GoB results show that to keep more than 77% of participants

Table 5.26: Percentage of participants reporting disturbances.

Test case	Blurriness	Blockiness	Freezes once	Freezes several times
TC1	51.85%	14.81%	7.4%	0%
TC2	55.55%	22.22%	14.81%	0%
TC3	37.04%	25.93%	7.4%	11.11%
TC4	37.04%	18.52%	3.71%	7.41%
TC2’	59.26%	14.81%	0%	3.71%
TC3’	29.63%	18.52%	3.71%	7.41%

Table 5.27: Mean values of video impairments and rated video quality.

Test case	**Blurriness mean/median/StDev**	**Blockiness mean/median/StDev**	**MOS VQ**	**%GoB VQ**
TC1	6.26/6.22/0.75	32.32/32.12/2.62	3.67	62.96%
TC2	6.37/6.25/0.59	31.11/30.74/2.53	3.41	59.26%
TC3	6.82/6.78/0.78	32.66/32.22/2.9	3.93	77.78%
TC4	6.79/6.85/0.72	31.63/31.28/2.79	3.59	62.96%
TC2’	6.49/6.25/0.61	31.54/31.15/2.55	3.44	62.96%
TC3’	6.98/6.96/0.67	32.37/32.17/2.16	4	81.48%

satisfied (rating video call with 4 or 5) scenarios where VQ MOS was above 3.93 should be considered, meaning that resolution should be at least 320x430 px, with 300 kbps video bitrate and frame rate of 20 fps. In contrast to US5, here we can not observe significant drop - the worst scenario yielded 3.41 VQ MOS, at which level 59.26% of the participants were satisfied. The reason likely lies in the fact that in US6 participants did not rate the scenario with 120x180 px video resolution, which we found to be unacceptable in terms of user perceived quality in study US5.

We divided test cases based on the frame rate values 10 fps (TC1) and 20 fps (TC2 and TC2', TC3 and TC3', TC4) into two groups. As expected, the test case with 10 fps had on average the blurriest image, but TC1 did not scored the lowest average video quality rating. Looking at the 20 fps test cases, a positive trend can be observed, meaning that higher objective video quality in terms of blurriness and blockiness did yield higher MOS. Although we can identify some correlation, in interactive services participants (especially younger ones) can introduce a lot of movement during the communication and subsequently cause additional inconsistent blurring. The following histograms show the blurriness and blockiness values from all test scenarios associated with corresponding video quality ratings (Figure 5.21, 5.20). Similar to the results of our previous study, histograms for perceived video quality (VQ) ratings have a notably different spread and correlated frequency of occurrences of values, since participants rated video quality most often as “Good”.

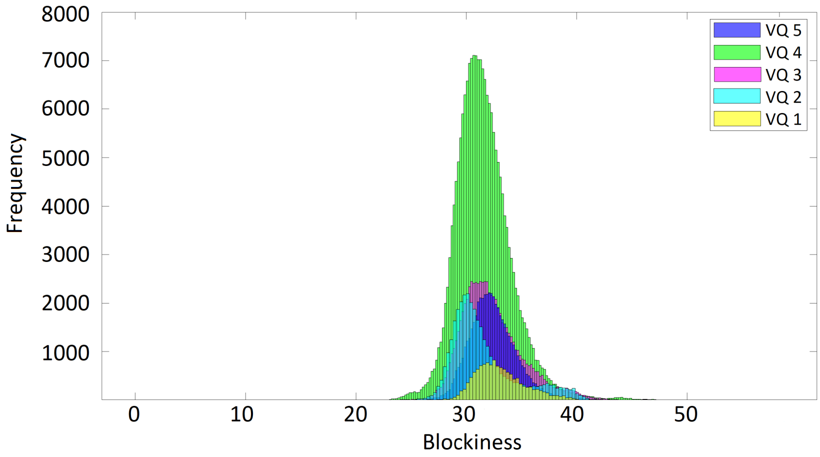


Figure 5.20: Frequency of blockiness values per frame for video quality (VQ) user ratings.

To obtain further insights into the potential of utilizing blurriness and blockiness as objective metrics in terms of estimating subjective multiparty video call quality, we analyzed the values between test cases with the same conditions, TC2 and TC2', as well as TC3 and TC3'. As illustrated on figures 5.22 and 5.23 histograms of blockiness and blurriness values are notably different, especially in case of blurriness with a plot that accentuates random artifacts in the data. Blockiness values were more evenly distributed, but still notably different. TC2' (VQ MOS 3.44) was less blocky and rated on average higher than TC2 (VQ MOS 3.41). In contrast to the

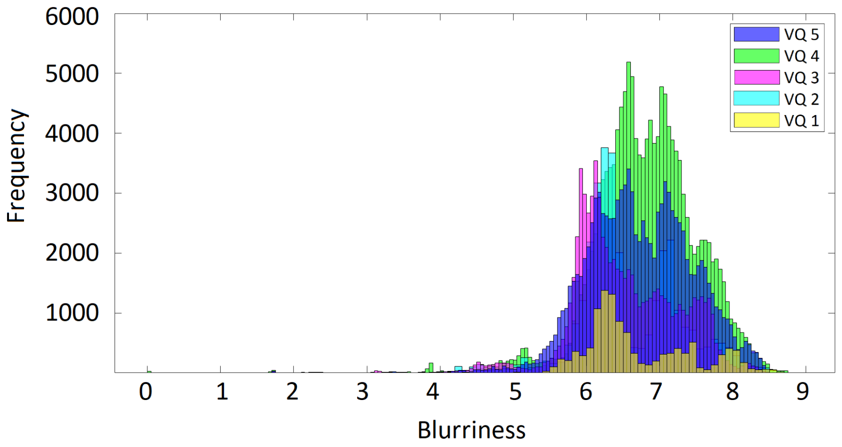


Figure 5.21: Frequency of blurriness values per frame for video quality (VQ) user ratings.

situation where TC3 as less blocky scored a lower mean rating (VQ MOS 3.93) as compared to TC3' (VQ MOS 4).

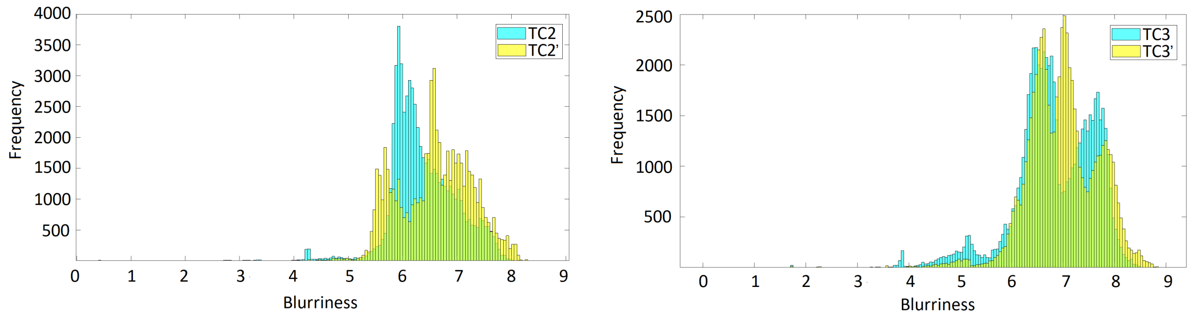


Figure 5.22: Frequency of blurriness values per frame for TC2-TC2' and TC3-TC3'.

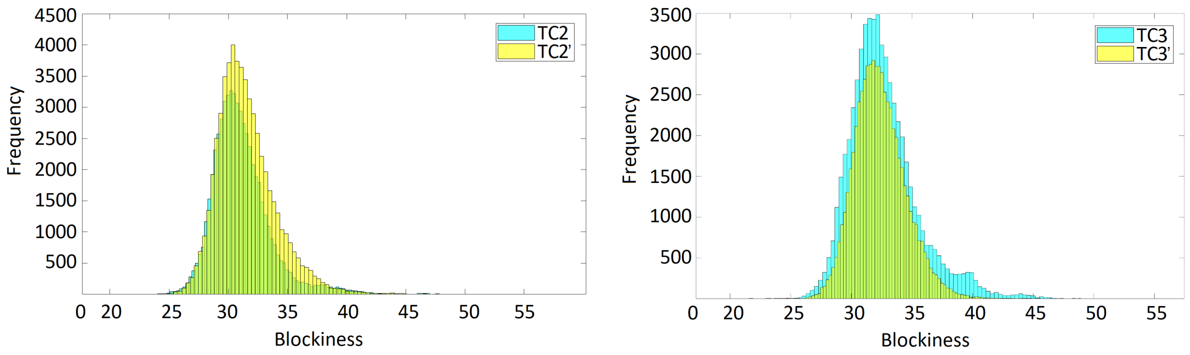


Figure 5.23: Frequency of blockiness values per frame for TC2-TC2' and TC3-TC3'.

Correlation between mean blurriness and blockiness values with objective metrics and subjective ratings

Pearson's correlation coefficient was used to measure how strong is the relationship between measured objective blurriness and blockiness and perceived video quality mean ratings and resolution (Table 5.28). The data show very strong positive correlation at the 0.01 and 0.05

Table 5.28: Correlations between mean and median blurriness and blockiness measured objective values and resolution and VQ MOS collected in US6.

Pearson correlation	Blurriness mean	Blurriness median	Blockiness mean	Blockiness median
Resolution	0.892*	0.854**	0.711	0.595
VQ MOS	0.705	0.655*	0.912*	0.869**

** Correlation is significant at the 0.01 level

* Correlation is significant at the 0.05 level

Table 5.29: Correlations between perceived impairments reported by participants (blurriness, blockiness, and freeze) and resolution and VQ MOS collected in US6.

Pearson correlation	Perceived blurriness	Perceived blockiness	Perceived freeze once	Perceived freezes several times
Resolution	-0.397	0.561	-0.170	0.866*
VQ MOS	0.002	0.358	-0.172	0.697

* Correlation is significant at the 0.05 level

level between perceived video quality MOS and blockiness mean and median value, respectively. Correlation between VQ MOS and blurriness median value was strong at 0.05 level with correlation factor 0.655. Both blurriness mean and median values showed very strong positive correlation with resolution at 0.05 and 0.01 level, respectively. Such results are in contrast to the computed coefficients in study US5, where blockiness showed strong positive correlation with resolution.

Taking into consideration perceived disturbances reported by participants (perceived blurriness, blockiness and freezes), results show significant positive correlation only between resolution and perceived freezes several times reported by participants, which can be attributed to the processing capabilities of the end user device (Table 5.29).

Perceived video quality model based on objective video metrics

Furthermore, we modeled the perceived video quality (PVQ) using predictors blockiness (BLO) and blurriness (BLU) (eq. 5.1). The predictor blockiness is statistically significant because its p-value is less than the significance level of 0.05, while the predictor blurriness is statistically insignificant with the p-value of 0.061.

$$PVQ = 0.266 \cdot BLO + 0.28 \cdot BLU - 6.5 \tag{5.1}$$

The PVQ regression model was significant with p-value less than 0.05. The coefficient of

Table 5.30: Model summary.

Model	R	R^2	Adjusted R^2	Std. error of the estimates
QoE	0.957[a]	0.976	0.953	0.0657

a. Predictors: (Constant), BLU, BLO

b. Dependent Variable: PVQ

determination R^2 indicates that 95.3% of the total variance is explained by the independent variables (Table 5.30). The accuracy of the PVQ estimation model is shown in Figure 5.24. Calculated data in Table 5.31 show that F is 43.179 of the variance generated by the regression, with degree of freedom (2, 8).

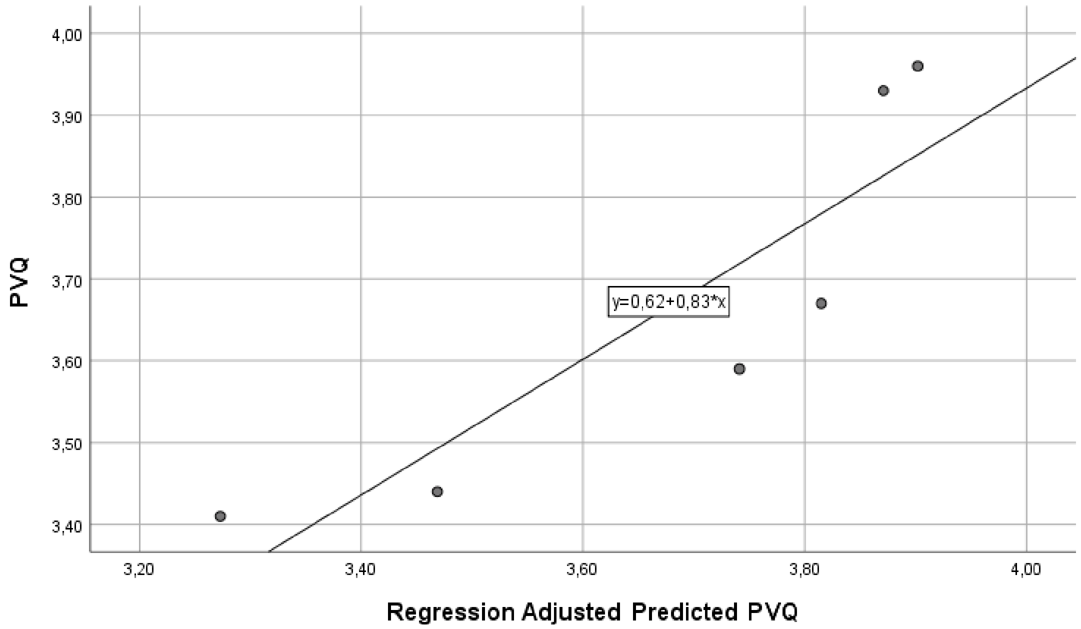


Figure 5.24: Accuracy of estimated video quality ratings (horizontal axis) compared to the actual video quality ratings (vertical axis) collected in user study US6 (based on the model given in Equation 5.1).

Participants noticed and reported noticing blurred image during sessions more often than a blocky image. Thus, this data could be used to improve users QoE in a way if there are enough available resources, bitrate could be increased. On the other hand, blockiness and blurriness could be hard to distinguish and perceive due to the small preview screen size and interaction, meaning that these objective video quality metrics in terms of multiparty audiovisual telemeeting on mobile devices are not the best metrics to use for perceived interactive quality evaluation. We note that in a dyadic call, where the situation with impairments is less complex, meaning that each participant can notice only the impaired video or audio of the other interlocutor, results could be different.

Table 5.31: QoE model variation analysis - ANOVA.

QoE Model	Sum of squares	df	Mean square	F	Sig.
Regression	2.66	2	0.133	30.69	0.01^b
Residual	0.013	3	0.004		
Total	0.279	5			

a. Dependent Variable: PVQ

b. Predictors: (Constant), BLU, BLO

Summary of key findings

The results in reported study US6 show that:

- In the context of three-party audiovisual calls established using smartphones with processing capabilities comparable to those tested (4GB of RAM), 300 kbps bitrate should be enough to transmit resolutions up to 320x430 px without impairment caused by CPU overuse or bitrate limitation.
- A resolution of 180x240 px per video in a multiparty context should be used in case of limited system or network resources.

5.5 QoE and PVQ estimation models derived from data collected in user studies US5 and US6

In this section, we report on two general types of models: 1) models designed to estimate overall QoE based on perceived video and audio quality, and 2) models to estimate PVQ based on video encoding parameters. We refrain from deriving overall QoE models using video coding parameters as predictors, but rather choose QoE features instead. This is due to the fact that QoE in the context of an audiovisual telemeeting is not only determined by perceived video quality, yet with perceived audio quality as well, along with additional features specific for interactive multimedia services, such as AV synchronization or ability to interact smoothly.

To build a model to estimate the value of QoE, we used data collected in studies US5 and US6. We use regression to develop a model that estimates values of the response variable based on the values of the predictors, perceived audio- and video quality features. It should be noted that some of the data sets were not well-modeled by a normal distribution and did show skewness and kurtosis, but ANOVA test is considered as robust against normality assumptions [149]. Also, dependent variable are measured at interval levels, since the difference between the points on the rating scale is assumed to be equal.

Table 5.32: QoE$_{gen}$ model variation analysis - ANOVA.

QoE$_{gen}$ Model	Sum of squares	df	Mean squares
Regression	139.228	4	34.807
Residual	0.053	7	0.008
Uncorrected total	139.281	10	
Corrected total	1.079	10	

R^2 = 1- (Residual Sum of Squares) / (Corrected Sum of Squares) = 0.95

5.5.1 QoE model for unimpaired sessions

The challenges of an unstable network environment (in terms of the packet loss or delay) make it difficult to estimate and model QoE accurately when sporadic disturbances have been present in the session. The duration of the disturbances, time to recovery, to which extent video quality has been degraded compared to the preset one, when did disturbances happen (at the beginning, middle or the end of the video call), was audio impaired, are all influence factors which might have significant impact on the perceived overall quality. Thus, we focused on the sessions with stable conditions in terms of network disturbances, and modeled QoE based on the results collected during the sessions with insignificant packet loss and delay, which we refer to as *unimpaired sessions*.

Perceived video quality (PVQ), perceived audio quality (PAQ) and overall quality ratings reported in user studies US5 and US6 served as input to derive a QoE model. Different forms of regression analysis were used in order to establish the relationship and model QoE. Results showed that a linear model, and model based on the generic audiovisual quality given in Equation 3.1 had the best fit. Based on the coefficients we obtain the following equation according to the generic model:

$$QoE_{gen} = 3.032 \cdot PAQ + 2.719 \cdot PVQ - 0.682 \cdot PAQ \cdot PVQ - 8.247 \quad (5.2)$$

The coefficient of determination R^2 indicates that 95% of the total variance is explained by the independent variables PAQ and PVQ (Table 5.32). We note however that due to the extrapolation, some combinations of audio and video quality ratings (such as PAQ=1 and PVQ=1) might result with the wrong estimation, considering these predictor values are outside the range ratings [2.44 - 4.14].

Based on the coefficients we obtain the Equation 5.3 with linear relationship:

$$QoE = 0.569 \cdot PAQ + 0.43 \cdot PVQ - 0.007 \quad (5.3)$$

Table 5.33: Model summary.

Model	R	R^2	Adjusted R^2	Std. error of the estimates
QoE	0.957[a]	0.915	0.894	0.10695

a. Predictors: (Constant), PAQ, PVQ

b. Dependent Variable: QoE

The QoE regression model was significant with $p < 0.001$. The R^2 indicates that 91.5% of the total variance is explained by the independent variables (Table 5.33). Both predictors, PAQ and PVQ have a statistically significant impact on the QoE because their p-values are less than the significance level of 0.05. The accuracy of the QoE estimation model is shown in Figure 5.25. Calculated data in Table 5.34 show that F is 43.179 of the variance generated by the regression, with degree of freedom (2, 8).

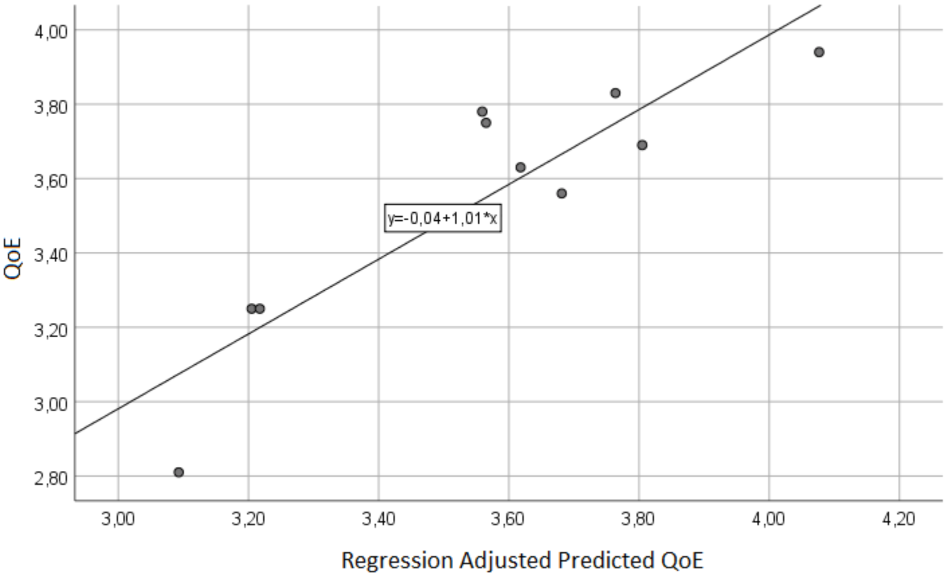


Figure 5.25: Accuracy of estimated QoE ratings (horizontal axis) compared to the actual QoE ratings (vertical axis) collected in user studies US5 and US6 (based on the model given in Equation 5.3).

Since the linear regression model is developed from the generic model (multiplication factor of independent variables equals zero), the generic model will always be either equally accurate or more accurate than the linear model. In this case, increasing the complexity of the expression to obtain more accurate interpolation is not significant. However, in certain cases the increased complexity does not justify the achieved accuracy of the model. Comparing linear QoE model to the QoE_{gen} model, both showed high R^2, but residual or the error sum of squares is lower in the QoE_{gen} model.

It is clear that for high quality telemeetings, providing good video and audio quality is of great importance. However, models also showed that good audio quality can compensate for

Table 5.34: QoE model variation analysis - ANOVA.

QoE Model	Sum of squares	df	Mean square	F	Sig.
Regression	0.988	2	0.494	43.179	0.000[b]
Residual	0.092	8	0.011		
Total	1.079	10			

a. Dependent Variable: QoE

b. Predictors: (Constant), PAQ, PVQ

poor quality visuals. Hence, from the QoE perspective, in case of capacity constraints it is important to prioritize audio quality over video quality.

“Good or better” QoE metric

In addition to the MOS VQ values we look at the percentages of participants assessing the test condition as Good or Better (%GoB) referring to the ratio of participants rating 4 or 5. According to the reported results in US5 and US6 combined, Figure 5.26 visualizes plots of the %GoB percentage for overall- and video- quality MOS scores (where each point corresponds to a single test scenario, and objective video quality is arranged in ascending order). Results show to keep more than 65% of participants satisfied (rating video call with 4 or 5) scenarios where VQ MOS was above 3.58 should to be considered, meaning that resolution should be at least 240x360 px, with 300 kbps video bitrate and frame rate 15 fps. Significant drop occurs in case of test condition 120x180 px, 100 kbps, 15 fps, at which level only 4.17% of the participants were satisfied, rating the test condition with 2.33 VQ MOS.

5.5.2 Relationship between perceived video quality and video encoding parameters

To establish the relation between perceived video quality and video encoding parameters, we included video bitrate (VBR), resolution (R), and frame rate (FR) as predictors. Resolution is calculated as a multiplication of frame height and frame width divided by 1000, and video bitrate unit is kbps. We restricted the model to resolutions up to 360x480 px due to the frequent adaptation in case of a 480x640 px preset resolution. Based on the scatter plot and residual plot, we concluded that a linear model is not the best choice to model the data. Thus, we tried to fit the data using various polynomial, logarithmic, and rational models. Results showed that PVQ (Equation 5.4) can be modeled as follows:

$$PVQ = \frac{-116.723}{VBR} - \frac{15.775}{R} - 0.023 \cdot FR + 4.653 \tag{5.4}$$

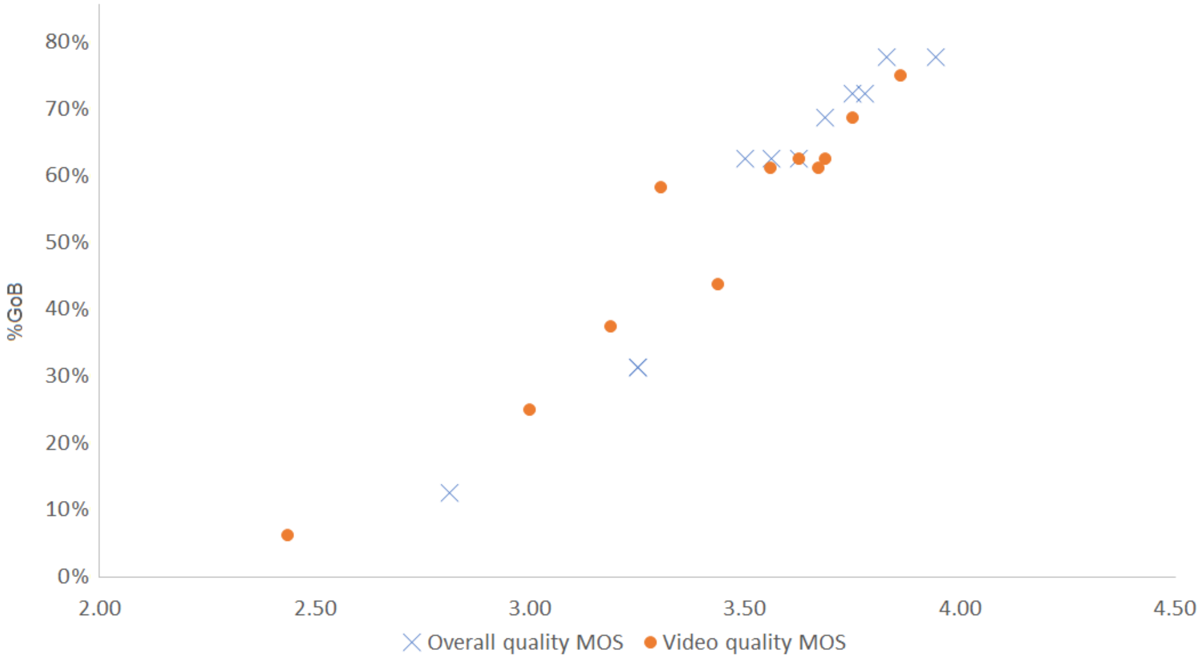


Figure 5.26: %GoB ratio of QoE and video quality MOS ratings collected in user study US5 and US6.

The R^2 indicates that 93.7% of the total variance is explained by the independent variables (Table 5.35).

Table 5.35: PVQ model (based on bitrate, resolution, and frame rate) variation analysis - ANOVA.

Source	Sum of squares	df	Mean square
Regression	129.586	4	32.396
Residual	0.107	7	0.015
Uncorrected total	129.692	11	
Corrected total	1.704	10	

R^2 = 1- (Residual Sum of Squares) / (Corrected Sum of Squares) = 0.937

Furthermore, we tried to enhance model accuracy by adding objective parameters blockiness and blurriness (mean values) as additional predictors (eq. 5.5).

$$PVQ = \frac{-118.372}{VBR} - \frac{15.56}{R} - 0.22 \cdot FR - 0.009 \cdot BLU + 0.02 \cdot BLO + 4.616 \tag{5.5}$$

Given our data, the model based on encoding parameters with additional predictors in terms of objective metrics (blurriness and blockiness) did not show any significant difference as compared to the model based on the encoding parameters only (R^2=0.938 calculated data in Table 5.36). However, further research is needed to determine whether additional objective video quality predictors can improve the estimation model for perceived video quality.

Table 5.36: PVQ model (based on bitrate, resolution, frame rate, blurriness and blockiness) variation analysis - ANOVA.

Source	Sum of squares	df	Mean square
Regression	129.586	6	21.598
Residual	0.106	5	0.021
Uncorrected total	129.692	11	
Corrected total	1.704	10	

R^2 = 1- (Residual Sum of Squares) / (Corrected Sum of Squares) = 0.938

Table 5.37: Blurriness model (based on bitrate, resolution and frame rate) variation analysis - ANOVA.

Source	Sum of squares	df	Mean square
Regression	389.508	3	129.836
Residual	0.098	6	0.016
Uncorrected total	389.607	9	
Corrected total	0.395	8	

R^2 = 1- (Residual Sum of Squares) / (Corrected Sum of Squares) = 0.751

5.5.3 Relationship between objective video quality metrics and video encoding parameters

To establish the relation between objective video quality metrics and video encoding parameters, we used the ratio of a bitrate and multiplication of resolution and frame rate, where resolution (multiplication of frame height and frame width was divided by one thousand). We used mean values of blockiness and blurriness to build the model. With all measurements included, we could not yield a higher accuracy model more than 40%. We identified and removed two outliers (TC1 and TC2 from US5) from a data sample collected in user studies US5 and US6. After data exclusion, by analyzing accuracy of fit for multiple different non-linear models, we found that a model with exponential function provided the highest accuracy (eq. 5.6):

$$Blurriness = 0.861 \cdot exp(-3.706 \cdot \frac{VBR}{R \cdot FR}) + 6.225 \tag{5.6}$$

Taken as a set, the video encoding parameters (bitrate [kbps], resolution and frame rate) account for 75.1% of the variance in blurriness (Table 5.37).

To establish the relation between blockiness and video encoding parameters we used ratio of a bitrate and multiplication of resolution and frame rate, where resolution (multiplication of frame height and frame width was divided by thousand). We found that a quadratic polynomial

Table 5.38: Blockiness model (based on bitrate, resolution and frame rate) variation analysis - ANOVA.

Source	Sum of squares	df	Mean square
Regression	10236.856	4	12559.214
Residual	1.803	3	0.601
Uncorrected total	10238.659	7	
Corrected total	11.601	6	

R^2 = 1- (Residual Sum of Squares) / (Corrected Sum of Squares) = 0.845

function models the dependency with highest accuracy (eq. 5.7).

$$Blockiness = 18.044 \cdot (\frac{VBR}{R \cdot FR})^2 - 19.621 \cdot \frac{VBR}{R \cdot FR} + 41.57 \tag{5.7}$$

Taken as a group, the video encoding parameters (bitrate [kbps], resolution and frame rate) account for 84.5% of the variance in blockiness (Table 5.38).

Taking into consideration video encoding parameters as a predictors, both, blurriness and blockiness model showed better conformation to the non-linear function, exponential and quadratic respectively. Models also showed significantly higher accuracy when an independent variable was included in the form of a ratio *Bitrate/(Resolution·Frame rate)*, confirming the fact that video artifacts become more perceptible in case of insufficient bitrate. Therefore, one of the key challenge is to find the right amount of video bitrate (which will enable acceptable QoE) without being too generous and waste precious resources. In the section 7.1, we describe how to determine a sufficient video bitrate for specific resolution and frame rate for multiparty audiovisual telemeetings on mobile devices.

However we note that derived models are based on the collected results combined from US5 and US6. In case of modeling objective metrics using results only from US5 or US6, models would not be same. The reason for that lies in the different correlations of blurriness and blockiness with resolution and perceived video quality in US5 and US6. In US5, blockiness showed a very strong positive correlation with resolution at the 0.01 significance level, while in US6 blurriness showed very strong positive correlation with video resolution at the same significance level. Thus, to gain more confidence in given results, blurriness and blockiness reference values corresponding to the specific video quality level should be established. We allowed participants to act freely, so as to obtain user opinions in the natural context mimicking as much as possible real life-scenarios. It was, though, a benefit that comes at a cost when trying to involve blurriness and blockiness into the QoE equation. For future studies involving measuring artifacts, we recommend to introduce limitations into the experiments, concerning participant movement, such as mandatory use of stands for smartphones, and sitting on fixed chairs (without possibility to swivel). In such way we would avoid at least to some extent fluctuating values introduced by participants and obtain more stable results for blockiness, and blurriness especially.

Summary of key findings

- Low motion video does not require high frame rates. Thus, varying the frame rates did not contribute much to the perceived quality. Additionally considering that each participant's stream is previewed in a small window on the smartphone display, it becomes clear why frame rate was statistically insignificant.
- An estimation model for objective video quality metrics (blurriness and blockiness) based on video encoding parameters showed non-linear dependency and highest accuracy when using the ratio of bitrate to resolution and frame rate multiplication as an independent variable.

5.6 Validation of proposed models

In this section we validate the proposed regression models (for multiparty audiovisual telemeetings on mobile devices) for estimation of QoE and perceived video quality. The aim of the validation process is to test whether results of the regression analysis on the sample can be extended to another chosen sample. The models are designed to quantify the relationship between QoE and perceived audio and video quality, where perceived video quality is based on the video encoding parameters (bitrate, resolution, frame rate).

5.6.1 Validation results analysis

With the goal being to assess and measure how effectively the models describe the outcome variable, the use of independent data to fit and test the model is preferred when building the model to estimate the dependency for future subjects. Thus, to indicate how well a model will perform, the following procedures are useful in checking the validity in case of regression modeling [150]:

- comparison of the model estimations with theoretical models and simulation results,
- collection of new data to validate model estimations,
- reservation of a portion of the available data.

Given the fact that this research presents a first attempt of QoE modeling and its features in the context of multiparty video calls on mobile devices, a comparison with previous research results was not possible. Thus, to obtain an independent measure of the model estimation accuracy, we relied on the reservation of a portion of the available data procedure, known as the data splitting method. According to the procedure, we differ two samples from the same population, namely the training sample (used for fitting) and validation sample (used to examine model efficiency). An important characteristic of the samples is that they belong to and represent the same population, while being distinct and independent from one another. To build and fit the models given in Equations 5.3 and 5.4, we used a training sample. In this section, we evaluate the performance of the estimation model by applying the model to a validation sample. Splitting the data sets will avoid usage of the same sample twice. If the estimation model is applied to a population including already used samples when building the model, the performance measure used to fit the model may be biased in favor of the model. The estimation model can perform in an optimistic way on the training sample and can show lower performance on the validation sample. For the training data set, we focus on evaluating how well the model fits the data used to build the model. For the validation data set, the aim is to measure how accurately the estimation model estimates the outcome variable on an unseen sample.

The sample data set size has to be determined so as to provide an adequate amount of data in the training data set to build the model, and a corresponding amount of data in the validation data set to successfully validate the model. Our original population (referring to data collected in the scope of user studies US5 and US6) had a size of 330 reported ratings per rated category: perceived audio quality, perceived video quality, and overall quality. We randomly allocated 67% of the population data for the training sample, while the remaining 33% was used as the validation sample. Each split divided the population into the two subsets, with the training data set containing 220 reported ratings and the validation data set containing 110 reported ratings. The training and validation set had an approximately equivalent gender distribution per each set. A statistical analysis was performed on both, training and validation data set.

As already mentioned, our goal is to evaluate and measure how effectively the models estimate the outcome variable, both on the training data set and validation data set. Methods used for assessing the measure of model performance include Pearson correlation coefficient, MAD (Mean Absolute Deviation, eq. 5.8), RMSE (Root Mean Square Error, eq. 5.9) and MAPE (Mean Absolute Percentage Error, eq. 5.10).

$$MAD = \frac{\sum |Error|}{n} \tag{5.8}$$

$$RMSE = \sqrt{\frac{\sum (Error)^2}{n}} \tag{5.9}$$

$$MAPE = \frac{\sum |Error/Validation_{PVQMOS}|}{n} \tag{5.10}$$

Similarity measures, such as the Pearson coefficient of correlation measures in a dimensionless way the linear relationship between two variables, and the interpretation of the coefficient

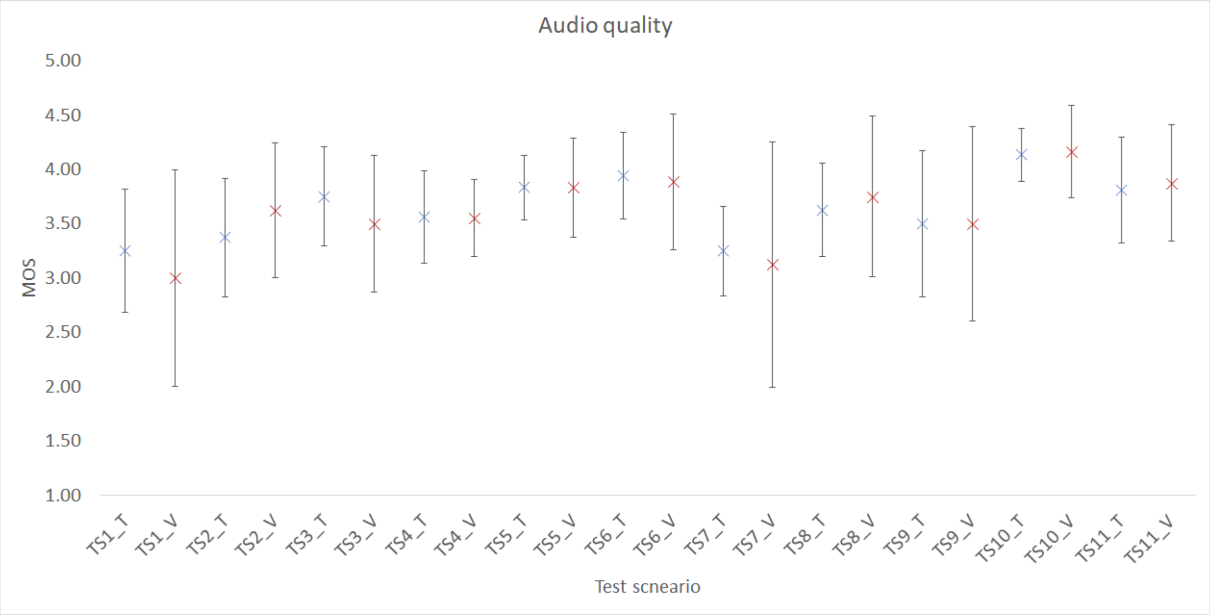


Figure 5.27: Comparison of audio quality MOS between training and validation data sets (95% confidence intervals shown).

is straightforward, whereas the change of magnitude in one variable may be correlated or unrelated with the magnitude of another variable.

5.6.2 Assessment of model performance

Observations are similar for both data sets and both models (QoE and PVQ). Correlation results are high, and the ratings are not over- or underestimated systematically. In addition, for each rated quality category, the ratings reported in training and validation set are compared based on the 95% confidence intervals (Figures 5.27, 5.28, 5.29). We use *TSx_T* to denote values obtained from the training set and *TSx_V* to denote values from the validation set for each specific test scenario *x*.

In Table 5.39 and Table 5.40, we portray the obtained values when considering models for estimating QoE based on PAQ (perceived audio quality) and PVQ (perceived video quality) features, and perceived video quality based on video encoding IFs (bitrate, resolution, and frame rate). The last column in the tables (Difference) shows the difference between the actual and estimated quality for the training data set. We once again note that the derived models are as follows (Equations 5.3 and 5.4):

$$QoE = 0.569 \cdot PAQ + 0.43 \cdot PVQ - 0.007$$

$$PVQ = \frac{-116.723}{VBR} - \frac{15.775}{R} - 0.023 \cdot FR + 4.653$$

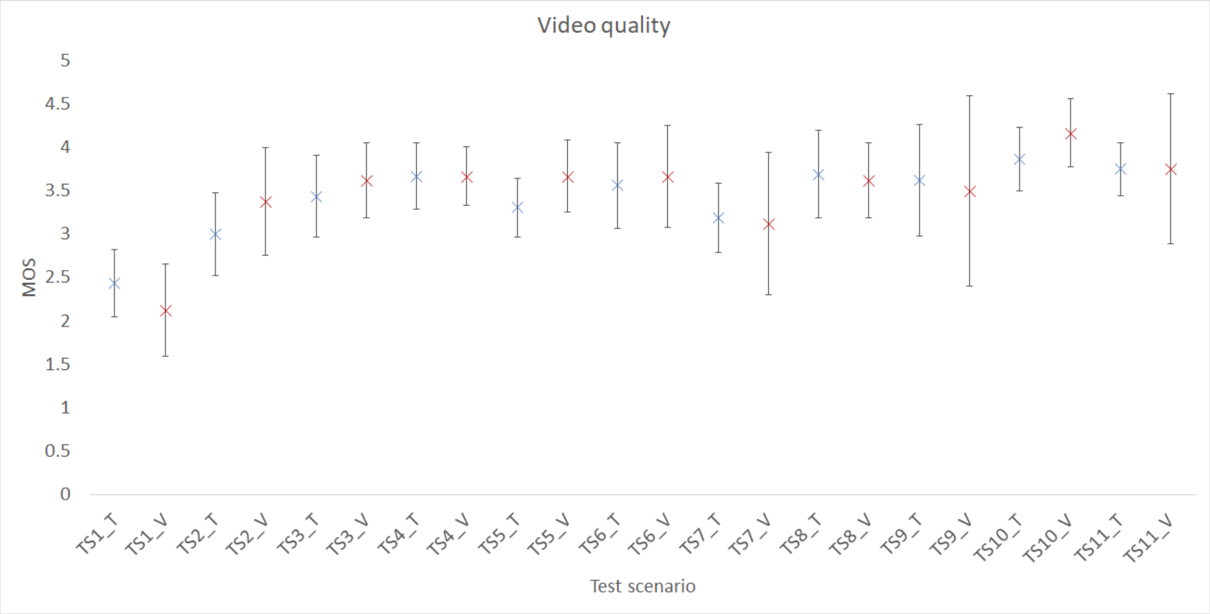


Figure 5.28: Comparison of video quality MOS between training and validation data sets (95% confidence intervals shown).

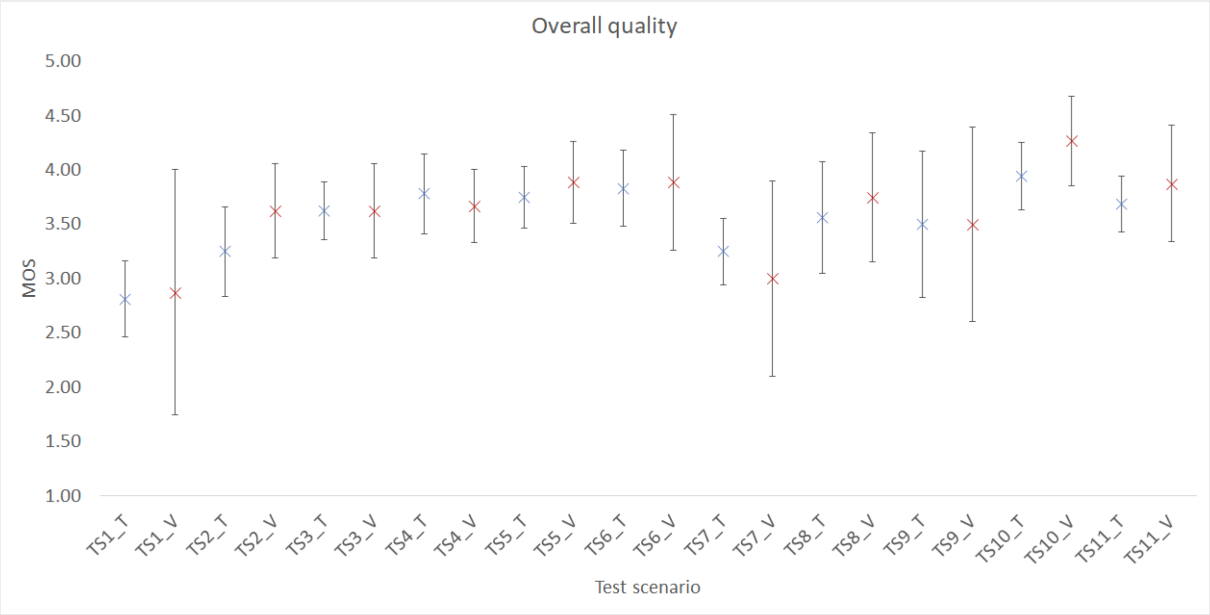


Figure 5.29: Comparison of overall quality MOS between training and validation data sets (95% confidence intervals shown).

Table 5.39: Results of validation of the QoE model.

Test scenario	Training data set QoE	Validation data set QoE	Estimated QoE for validation data set	Difference |Error|
TS1: 120x180 px, 100 kbps, 15 fps	2.81	2.88	2.63	0.18
TS2: 120x180 px, 200 kbps, 20 fps	3.25	3.63	3.53	0.28
TS3: 180x240 px, 200 kbps, 15 fps	3.63	3.63	3.56	0.07
TS4: 180x240 px, 300 kbps, 10 fps	3.78	3.67	3.61	0.17
TS5: 180x240 px, 300 kbps, 20 fps	3.75	3.89	3.76	0.01
TS6: 240x320 px, 300 kbps, 20 fps	3.83	3.89	3.80	0.03
TS7: 240x360 px, 150 kbps, 15 fps	3.25	3.00	3.13	0.12
TS8: 240x360 px, 200 kbps, 15 fps	3.56	3.75	3.70	0.14
TS9: 240x360 px, 300 kbps, 20 fps	3.50	3.50	3.50	0.00
TS10: 320x430 px, 300 kbps, 20 fps	3.94	4.28	4.17	0.23
TS11: 360x480 px, 300 kbps, 15 fps	3.69	3.88	3.83	0.14

Table 5.40: Results of validation of the PVQ model.

Test scenario	Training data set PVQ MOS	Validation data set PVQ MOS	Estimated PVQ MOS for validation data set	Difference |Error|
TS1: 120x180 px, 100 kbps, 15 fps	2.31	2.38	2.37	0.06
TS2: 120x180 px, 200 kbps, 20 fps	3.19	3.09	3	0.19
TS3: 180x240 px, 200 kbps, 15 fps	3.5	3.39	3.5	0.03
TS4: 180x240 px, 300 kbps, 10 fps	3.69	3.71	3.56	0.09
TS5: 180x240 px, 300 kbps, 20 fps	3.5	3.58	3.39	0
TS6: 240x320 px, 300 kbps, 20 fps	3.46	3.68	3.55	0.11
TS7: 240x360 px, 150 kbps, 15 fps	3.19	3.24	3.16	0.06
TS8: 240x360 px, 200 kbps, 15 fps	3.69	3.5	3.63	0.13
TS9: 240x360 px, 300 kbps, 20 fps	3.56	3.69	3.62	0.06
TS10: 320x430 px, 300 kbps, 20 fps	4.06	3.74	4.11	0.18
TS11: 360x480 px, 300 kbps, 15 fps	3.69	3.81	3.87	0.05

Table 5.41: Performance measures for QoE and PVQ estimation models, based on validation using a test data set.

Model	R^2	Pearson correlation	MAD	RMSE	MAPE
QoE	0.915	0.928	0.12	0.15	3.61%
PVQ	0.937	0.972	0.09	0.10	2.56%

We fitted both models on unseen (validation) data. The means of the estimated and actual quality values appear to be strongly correlated for both QoE and PVQ model (Pearson's correlation coefficient of 0.928, respectively 0.972). Both models performed well, where the PVQ model performed slightly better in terms of estimation error than the QoE model. Calculated MAD shows that the deviation of estimated value from the actual one is 0.09 for PVQ and 0.12 for QoE (Table 5.41). RMSE shows that the PVQ model could be off by 0.10, and 0.15 for QoE model. The model estimations are off by 3.61% on average in case of the QoE model, while for the PVQ model estimations are off by 2.56%. Considering the number of impact factors, especially in the multiparty mobile context, we conclude that both models provide a good level of estimation accuracy.

For visualization purpose, Figures 5.30 and 5.31 show the plot of the actual training and validation set of MOS scores, as well as estimated MOS for the validation data. Visual inspection shows that the QoE and PVQ models slightly underestimate mean ratings in test cases around a resolution 180x240 px (TS3, TS4, and TS5). In case of the other test scenarios, it can be observed that the estimated quality ratings are well modeled.

We note that the proposed models have limitations imposed by the experimental setup aiming to address specific influence factors (such as encoding video parameters) and QoE features. Despite these limitations, we were able to draw conclusions on the degrees to which chosen IFs and features impacted perceived video quality and QoE. However, to improve the accuracy of the proposed multidimensional models in the future, further analyses are needed aiming to include additional dimensions representing the perceptual quality space for multiparty telemeetings on mobile devices, such as perceived interactivity or audio-video synchronization.

5.7 Chapter summary

In this chapter, we presented subjective studies designed to investigate the impact of end user device hardware and video encoding parameters, such as bitrate, resolution and frame rate, on perceived video quality and overall user experience.

Throughout the course of our research studies, we noticed several issues which need to be considered. First of all, we noticed that during the course of conducted test sessions, some par-

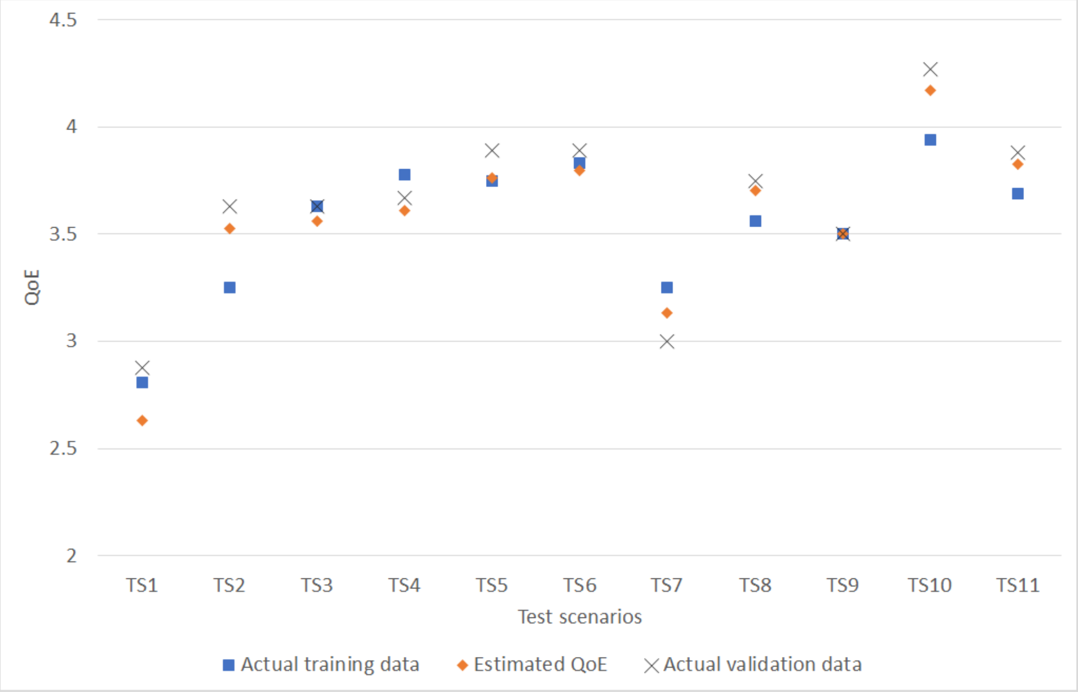


Figure 5.30: Comparison of actual and estimated QoE mean ratings per test condition.

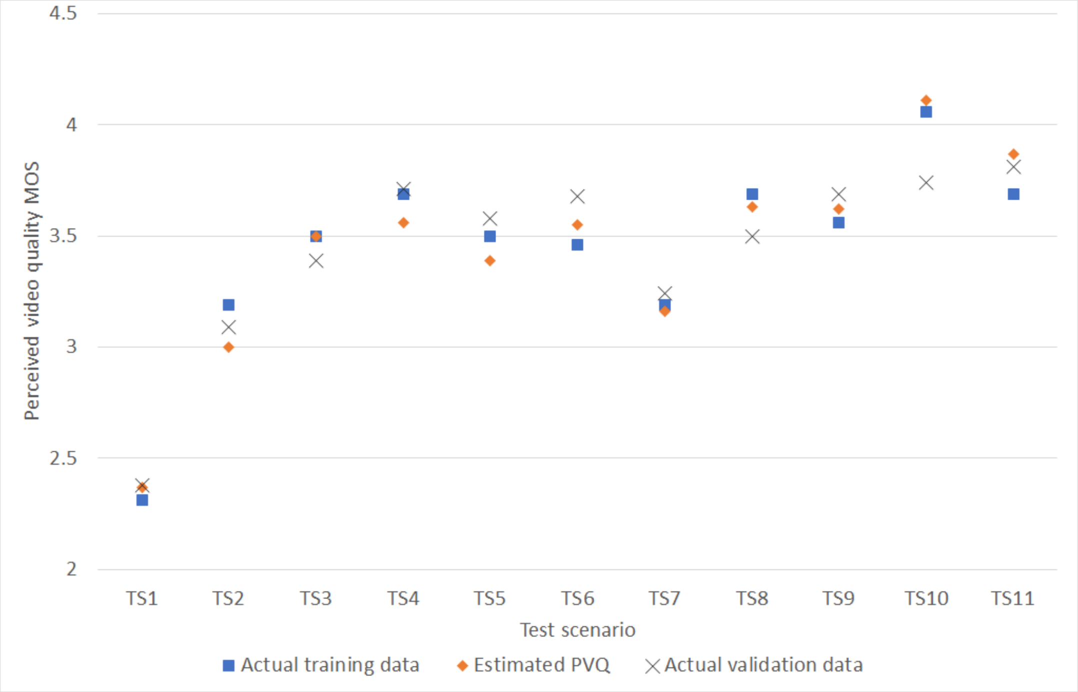


Figure 5.31: Comparison of actual and estimated perceived video quality mean ratings per test condition.

ticipants became more restless (even though the whole evaluation process lasted for a maximum of 45 minutes), thus causing additional movement and potential impact on both perceived quality and objective metrics. An increase in movement should be taken into consideration when defining target bitrates, as more dynamic scenes will likely require higher bitrates to achieve satisfactory QoE.

When testing on small screen sizes, especially in multiparty video calls where the preview window of each participant is relatively small, even if the objective video parameters are preset to significantly different values, quality assessment on a 5-pt. scale can produce similar results since it can be difficult to distinguish small perceived differences and relate them to the five ratings at the end of the session.

We further noticed that sometimes participants were not able to distinguish impairments, for instance reporting blockiness in cases when it was fairly low, instead of blurriness which was higher than average. This may possibly be attributed to the small preview size, short term and low strength of disturbances. Additionally, we noticed that participants engaged in the conversation can miss to detect short video freezes if: 1) the participant is not an active user, 2) audio quality is unimpaired, or 3) when the participant is staying still during the session.

Usually during a conversation, focus is on the active speaker. Hence, in a multiparty setup, the center of an eye gaze is commonly on the talking participant, while other participants in the group are outside the point of fixation. During our studies, all conducted in a leisure context, we noticed that occasional video impairments did not significantly impact overall perceived quality (however, we note that participants were only engaged in conversation, and were not focused on presenting to each other any particular visual cues). Thus, a key issue is to ensure enough resources to the active participant, prioritizing audio quality over video quality.

Finally, we considered multiple linear and nonlinear regression models in order to model overall QoE and perceived video quality PVQ. QoE was modeled in terms of dependent QoE features (perceived audio and video quality), while PVQ was modeled in terms of independent video encoding parameters (bitrate, resolution, frame rate), and in terms of video quality metrics blurriness and blockiness. The goal was to obtain findings with greater accuracy by analyzing how each of the predictors contribute to the perceived video quality and overall QoE. A summary of regression models derived from data in user studies are shown in Figure 5.32.

We described the procedure related to evaluation and validation of proposed multidimensional models for QoE and PVQ for three-party audiovisual telemeetings on mobile end user devices held in a leisure context. We measured estimation model efficiency with two independent data sets, training and validation, obtained from the population, by reservation of a portion of the collected data. Results have been analyzed and showed that both proposed models provide good accuracy and can be used for QoE and perceived video quality estimation as a QoE feature in terms of multiparty video calls.

Having addressed system factors, in the following chapter we will address subjective user studies designed to investigate the impact of network factors (such as packet loss and delay) and video encoding parameters (bitrate, resolution, and frame rate) on perceived video quality and overall user experience.

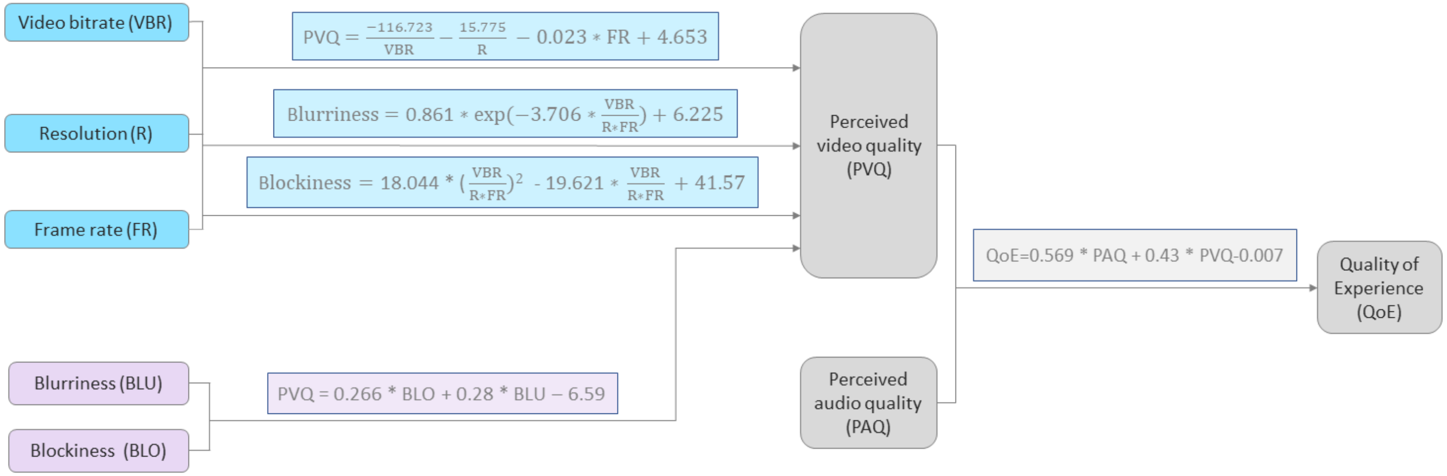

Video bitrate (VBR)
Resolution (R)
Frame rate (FR)
Blurriness (BLU)
Blockiness (BLO)
$PVQ = \frac{-116.723}{VBR} - \frac{15.775}{R} - 0.023 * FR + 4.653$
$Blurriness = 0.861 * \exp(-3.706 * \frac{VBR}{R*FR}) + 6.225$
$Blockiness = 18.044 * (\frac{VBR}{R*FR})^2 - 19.621 * \frac{VBR}{R*FR} + 41.57$
PVQ = 0.266 * BLO + 0.28 * BLU – 6.59
Perceived video quality (PVQ)
Perceived audio quality (PAQ)
QoE=0.569 * PAQ + 0.43 * PVQ-0.007
Quality of Experience (QoE)

Figure 5.32: Regression models based on the collected results in user studies US5 and US6.

Chapter 6

Impact of network factors on QoE for multiparty audiovisual telemeetings on mobile device

While previous expensive and complex desktop video conferencing solutions had a restricted reach, the emergence of the WebRTC paradigm has provided an opportunity to redefine the video communication landscape. In particular, technological advances in terms of high resolution displays and cameras have set the ground for multiparty video telemeeting solutions realized via mobile devices. A key challenge in the mobile context is managing quality in light of time varying network disturbances such as packet loss or delay. Hence, in this chapter, we present subjective user studies aiming to investigate network impairments on end user QoE. In study US3, we conducted an experimental investigation of the Google Congestion Control algorithm in light of packet loss and under various video encoding parameters, with the aim being to observe the impact of various adaptation scenarios on end user QoE (Section 6.1). Subsequently, study US4 served for comparison between test results (collected in US3) in a network impaired environment, with baseline results (collected in US4) in an unimpaired network environment with no inserted packet loss and application layer delay. (Section 6.2). Table 6.1 gives a brief overview and summarizes differences between studies US3 and US4.

Table 6.1: An overview of conducted subjective QoE studies.

User study	Participants, MIN/MAX/AVG age	End user device	Manipulated parameters
US3, 2017, [5]	16 males, 14 females, 1 fixed user per test group, 33/49/40	3 x Samsung S6	Video resolution, bitrate, frame rate, packet loss
US4, 2018, [41]	21 males, 6 females, 20/29/21	3 x Samsung S6	Video resolution, bitrate, frame rate

6.1 User study US3 - Impact of network factors

To enable WebRTC applications to load quicker and run smoother, the Google Congestion Control algorithm (implemented in Google Chrome) has been designed to work with RTP/RTCP protocols and target real-time streams such as telephony and video conferencing. The GCC algorithm includes two control elements: a *delay-based* controller on the receiver side, and a *loss-based* controller on the sender side (which complements the delay-based controller if losses are detected). The congestion controller on the sender side bases decisions on measured round-trip time, packet loss, and available bandwidth estimates [151]. In short, if 2-10% of the packets have been lost since the previous report from the receiver, the sender rate will be kept unchanged. If more than 10% of the packets have been lost the rate will be decreased. If less than 2% of the packets have been lost, then the rate will be increased [152]. To explore how GCC handles network packet loss under different video resolution, bitrate, and frame rate constraints and how packet loss and delay impact perceived quality, we conducted an empirical study (results reported in [5]). This user study helps to answer the high-level research question RQ5 as defined in Figure 1.1.

6.1.1 Methodology

Experiments were conducted involving interactive three-party audiovisual conversations in a natural environment and leisure context over a Wi-Fi network with symmetric device conditions so as to eliminate the impact of different devices. Experiments were carried out in a controlled environment and used to collect subjective end user assessments, rating the impact of packet loss on perceived quality. Moreover, WebRTC call-related statistics were collected for the purpose of performance analysis.

The three-party video telemeeting was set up using a WebRTC application running on the Licode open source media server installed in a local network, to avoid impairments caused by a commercial network, enabling us to preconfigure application parameters: bitrate, fame rate, and video resolution (Figure 6.1).

These default settings were then dynamically adapted based on activation of the GCC algorithm in response to inserted loss. The Licode media server was installed on a laptop with Intel Core i5 Processor, 2.6 GHz, 8 GB RAM and Ubuntu 14.04. The LAN connection between end user devices and the media server was Wi-Fi 802.11, on port 3004. Experiments were conducted in a natural home environment, with all three participants taking part in the call using mobile phones Samsung Galaxy S6, Android ver.6.0.1 and Chrome 55.0.2883.91 (Figure 6.2).

We note that the participants were physically located in three separate rooms and could not see/hear each other outside of the established call. The rooms had the following dimensions LxWxH (cm): room 1 - 385x327x260, room 2 - 385x250x260, room 3 - 385x320x260.

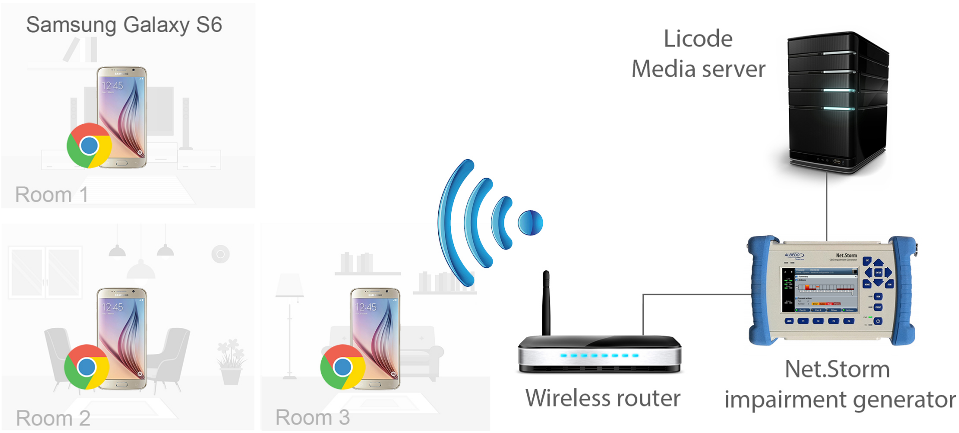


Figure 6.1: Testbed set-up over a LAN (user study US3).

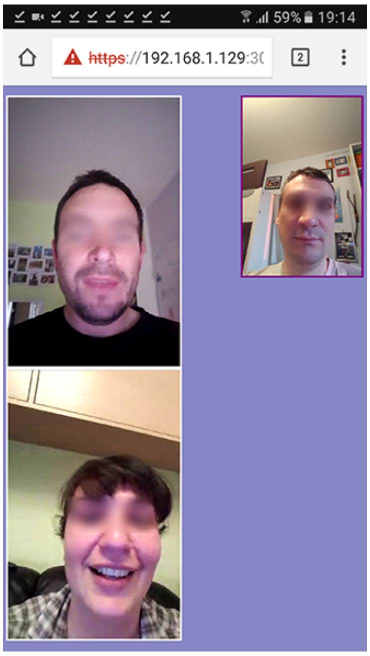


Figure 6.2: Example three-party video conversation in the Chrome browser. The upper right window portrays the local self-recording video.

Table 6.2: Test schedule used in user study US3.

Experiment	Video resolution [px]	Frame rate [fps]	Bitrate [kbps]
Test case 1 (TC1)	320x480	15	300
Test case 2 (TC2)	320x480	15	600
Test case 3 (TC3)	320x480	20	300
Test case 4 (TC4)	320x480	20	600
Test case 5 (TC5)	480x640	15	300
Test case 6 (TC6)	480x640	15	600
Test case 7 (TC7)	480x640	20	300
Test case 8 (TC8)	480x640	20	600

Packet loss was artificially generated in the experiments using the Albedo *Net.Storm*[1] network emulator, which enabled frame loss insertion. Net.Storm is a hardware-based emulator with the capability to emulate different degradations or impairments in Ethernet / IP networks. We used the function *frame periodic burst* to drop frame bursts, with a configurable number of frames that make up each loss burst and the separation between loss bursts. Loss bursts were periodically inserted, with burst length of 10 frames, and burst separation of 5 frames between consecutive loss bursts. We initiated packet loss starting after the first minute of each test conversation, and lasting for 10 seconds, after which the impairment was turned off.

The test schedule consisted of participants rating 8 test conditions with different combinations of video resolutions (320x480 px and 480x640 px), bitrates (300 kbps and 600 kbps) and frame rates (15 fps and 20 fps), each lasting 3 minutes (Table 6.2). With 15 participant groups, overall 120 tests were performed.

A preliminary test was carried out to introduce participants with the test procedure and assessment questionnaire, but results were not taken into account. After each 3 minute session was finished, participants were asked to rate audio quality, visual quality, AV synchronization, and overall quality using a paper questionnaire and the five point ACR rating scale.

Thirty participants took part in the study, 16 male and 14 female subjects, with an average age of 40 years (min 33, max 49). Participants were divided into 15 groups, with one fixed user added to each group as a third participant, to monitor the service and help keep the conversation flowing (this fixed third participant did not provide any subjective ratings). All participants were employed, 9 of them with high school education and 21 with a University degree. Participants reported having previous experience with the following video conversation applications (numbers indicate no. of participants): Skype (23), Viber (15), WhatsApp (13), Google hangouts (4), Facebook (1). The Croatian language was chosen to represent a natural interactive free

[1] http://www.albedotelecom.com/pages/fieldtools/src/netstorm.php

conversation, without any specific preassigned tasks. The selected subjects were not experts in audiovisual communications. Sixteen subjects have previously participated in subjective assessment. Subjects were volunteers, all with normal hearing, and 16 of them have corrected vision.

6.1.2 Results

Overall test statistics obtained from *webrtc-internals* data across all test sessions are given in Table 6.3. The lowest recorded resolution was 160x240, with frame rate 1 fps, and bitrate 15 kbps. In some cases, bitrates with values around 30 kbps lasted for approximately 30 seconds, which is a significant period in the context of 3 minute-long conversations. In average, TC1 managed to maintain preconfigured video encoding values for the longest time during the session. The default resolution of 320x480 px occurred during the conversation in 76.33% of session time. The default frame rate of 15 fps was observed in 73.61% of overall session time. TC6 maintained a default resolution of 480x640 px for only 21.77% of session time. In TC7, the default frame rate of 20 fps showed up with the lowest frequency in 46.37% of session time.

We found that all test conditions provided on average at least “Fair” audio, video, and overall quality, as well as AV synchronization. TC1 provided the highest average rating for audio quality (3.47) with the following settings for all flows: 320x480 px resolution, 15 fps, and 300 kbps encoding bitrate. The highest synchronization ratings (3.63) were provided by TC6, with 480x640 px resolution, 600 kbps, and 15 fps. TC6 also received the highest average score for overall quality (3.6), while TC8 received the highest mean rating for video quality (3.63). Three out of four highest rated categories belong to the 480x640 px resolution setup, which was ultimately reduced to the 360x480 px or even lower (to the 240x320 px in case of a 300 kbps), showing that lower objective video quality can be utilized by future service adaptation strategies in terms of setting thresholds for video encoding parameters.

To provide better insights into rating distributions, Fig. 6.3 shows the percentage of participants providing each rating score for audio quality, video quality, AV synchronization and overall quality for each test condition. In TC1, TC4, and TC6, more than 50% of participants rated audio quality as “Good” or higher. While in TC1 60% of participants rated AV synchronization at least “Good”.

In test case TC8, more than 56% of participants rated video quality as “Good” or “Excellent”. In case of overall quality and test case TC6, more than 63% of participants rated it as “Good” or higher. On the other hand, TC1 with the lowest preset objective video quality was the only test case where rating “Bad” was never given in any rating category. Thus, lower objective video quality can be a trade-off inherent in multiparty video call in terms of limited resources.

We used a one way ANOVA to check for significant differences between audio quality, video quality, AV synchronization and overall quality for each test condition. Results given in

Table 6.3: WebRTC internals collected and analyzed data of mean values per test condition.

Test case	15 fps 300 kbps	15 fps 600 kbps	20 fps 300 kbps	20 fps 600 kbps
Percentage of session time where actual streamed resolution corresponded to the set 320x480 px resolution	76.33%	76.22%	73.06%	73.95%
Percentage of session time where actual streamed resolution corresponded to decreased 240x360 px resolution	13.83%	13.01%	13.11%	16.52%
Percentage of session time where actual streamed resolution corresponded to decreased 160x240 px resolution	9.84%	10.54%	13.65%	9.31%
Default frame rate	73.61%	63.62%	50.22%	49.73%
≥ 13 fps	93.74%	82.48%	92.27%	89.66%
6-13 fps	4.25%	3.87%	1.32%	1.87%
1-6 fps	1.77 %	0.46%	0.18%	0.5%

Test case	15 fps 300 kbps	15 fps 600 kbps	20 fps 300 kbps	20 fps 600 kbps
Percentage of session time where actual streamed resolution corresponded to the set 480x640 px resolution	32.05%	21.77%	48.05%	29.23%
Percentage of session time where actual streamed resolution corresponded to the decreased 360x480 px resolution	47.45%	74.09%	26.21%	65.49%
Percentage of session time where actual streamed resolution corresponded to the decreased 240x320 px resolution	17.01%	2.48%	24.95%	3.58%
Percentage of session time where actual streamed resolution corresponded to the decreased 180x240 px resolution	1.26%	1.66%	0.78%	1.68%
1-6 fps	0.77%	0.83%	0.4%	0.66%
6-13 fps	3.24%	3.51%	1.6%	1.61%
≥ 13 fps	89.81%	89.24%	95.27%	93.86%
default frame rate	67.95%	69.05%	46.37%	48.67%

Table 6.4: Highest MOS values.

Test conditions	Evaluated	MOS ratings
TC1 320x480 px, 15 fps, 300 kbps	Audio quality	3.47
TC8 480x640 px, 20 fps, 600 kbps	Video quality	3.63
TC6 480x640 px, 15 fps, 600 kbps	AV synchronization	3.63
TC6 480x640 px, 15 fps, 600 kbps	Overall quality	3.6

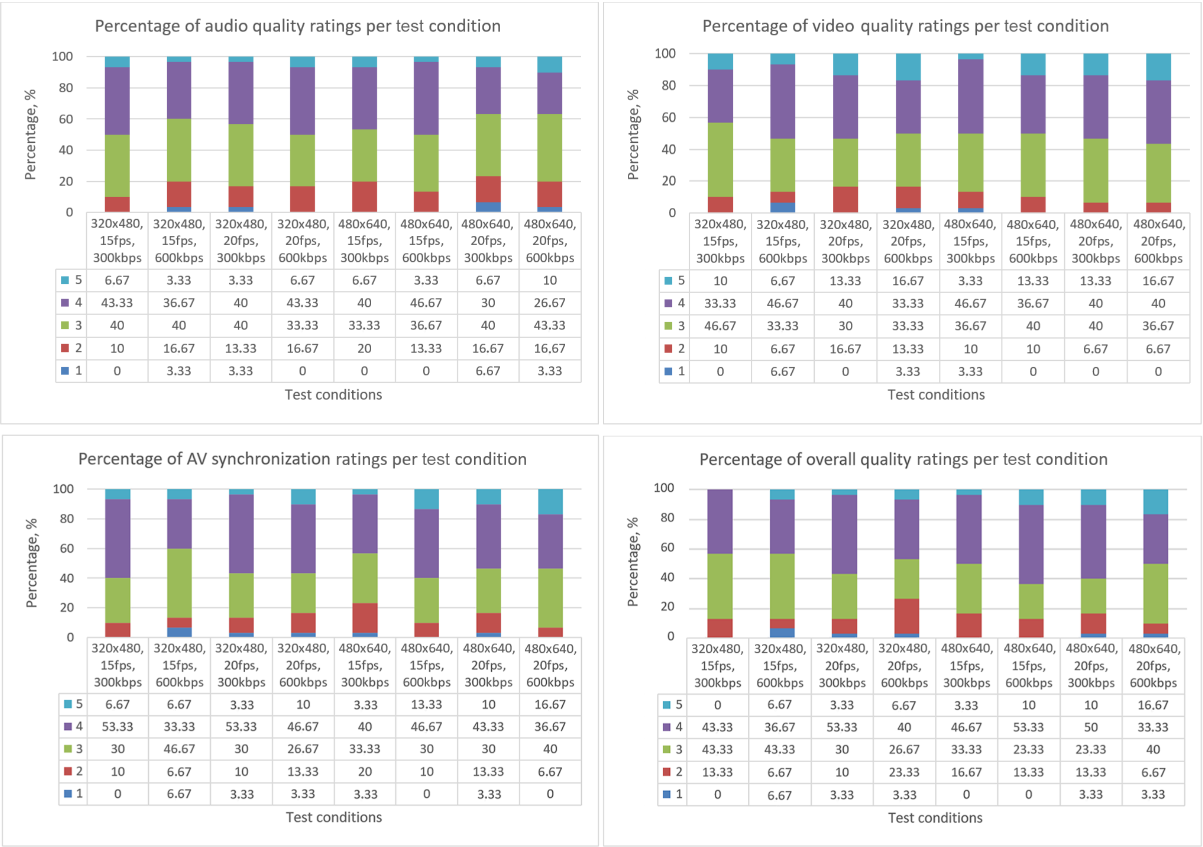


Figure 6.3: Distribution of ratings per test condition for audio quality, video quality, AV synchronization and overall quality.

Table 6.5: ANOVA analysis results for audio quality, video quality, AV synchronization and overall quality per each test condition.

Test case	SS	df	MS	F	P-value	F crit
320x480 px 15 fps 300 kbps	1.09	3.00	0.36	0.62	0.61	2.68
320x480 px 15 fps 600 kbps	0.63	3.00	0.21	0.24	0.87	2.68
320x480 px 20 fps 300 kbps	0.89	3.00	0.30	0.38	0.77	2.68
320x480 px 20 fps 600 kbps	1.09	3.00	0.36	0.39	0.76	2.68
480x640 px 15 fps 300 kbps	0.57	3.00	0.19	0.25	0.86	2.68
480x640 px 15 fps 600 kbps	0.96	3.00	0.32	0.46	0.71	2.68
480x640 px 20 fps 300 kbps	3.63	3.00	1.21	1.36	0.26	2.68
480x640 px 20 fps 600 kbps	3.50	3.00	1.17	1.40	0.25	2.68

Table 6.5 show that no significant difference exist between MOS scores.

Impact of inserted packet loss on performance:

In each test session participants reported service impairments. In response to inserted packet loss the GCC algorithm implemented within Chrome estimates new bandwidth values and subsequently invokes video quality reduction, including bitrate, resolution, and finally frame rate. As a result, the actual sent values start to differ from those initially configured. In case of the preconfigured resolution of 480x640 px, the adapted resolution preserved the aspect ratio when resized: 360x480 px, 240x320 px, and 180x240 px. In case of the 320x480 px preconfigured resolution, adapted actual streamed resolution corresponded to the decreased 240x360 px or 160x240 px resolution.

Ten seconds of inserted bursty packet loss caused 25 to 50 seconds of video conversation with lower quality, after which the service managed to restore values to those preconfigured on the media server. In some cases, the service never restored to the initial settings, but continued running on the reduced ones.

Subjects reported video loss of one participant after inserted packet loss in all test cases except the test case with lowest (TC1) and highest (TC8) preset objective quality. Complete video loss of one participant occurred in 8% of all sessions, with the video remaining lost until the end of the session. We note that this effect has also been observed and reported in previous work [153], where WebRTC is trying to adapt to the loss of link capacity but remains unrecovered after the network conditions were restored. While video loss had a significant impact on certain participants, for other participants it did not contribute to the quality perception, according to collected ratings. For example, in test group 8, Figure 6.4 portrays outgoing and incoming bitrates for TC2 (320x480 px, 15 fps, 600 kbps). What we observe is that quality degradation lasted for approximately 35 seconds. The video bitrate of one incoming participant stream

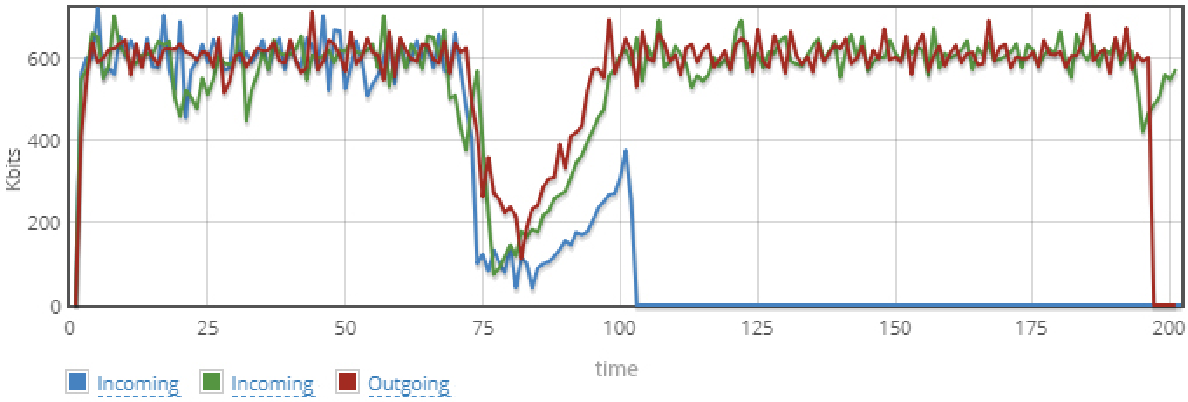


Figure 6.4: Example of incoming and outgoing bitrates for a session with a “lost” video stream (blue line) due to packet loss.

dropped to zero in the 100th second, and failed to recover for the remainder of the session. One participant in this case rated audio quality with “Poor” and video quality, AV synchronization and overall quality with “Bad”. Another participant from the same group rated audio quality with “Good” and video quality, AV synchronization and overall quality with “Fair”. Based on the results obtained in the study, we can conclude that in a leisure context, temporary loss of a video stream does not necessarily have a significant impact on QoE, as long as there is limited audio degradation.

We further discuss the influence of packet loss and delay on perceived audio and video quality for different test conditions (results shown in Table 6.6). Inserted packet loss exhibits similar patterns in all test conditions. The average number of lost packets for incoming audio streams ranged from 4.29% in TC3 to 5.17% in TC8, while for the video streams it ranged from 2.29% (T2) to 3.13% (TC8). Even though packet loss degraded perceived quality, the amount of degradation did not differ greatly between tested conditions. Packet loss also caused application layer delay, but we found no significant correlation between delay and packet loss average values. Average audio delay (on application layer) ranged from 328.85 ms (TC2) to 377.99 (TC1), and the video delay ranged from 306.79 ms (TC8) to 379.55 (TC5). In case of audio, the test case with highest average delay yielded the highest MOS score of 3.47 for perceived audio quality, which is contradictory to expectations. On the other hand, perceived video quality was rated with the highest MOS of 3.63 in case of the lowest delay. Results clearly show that besides packet loss and delay, other influence factors have great impact on the perceived quality in case of a multiparty video call.

Table 6.6: Average bitrate, delay, percentage of packet loss, and perceived media quality for audio and video streams across all test cases.

Test case	Stream	Bitrate [kbps]	Delay [ms]	Packet loss	Perceived quality
TC1	Audio	69.34	377.99	4.69%	3.47
	Video	278.36	344.79	2.66%	3.43
TC2	Audio	68.59	328.85	4.44%	3.2
	Video	504.03	373.33	2.29%	3.4
TC3	Audio	68.69	345.62	4.29%	3.27
	Video	275.33	318.67	2.39%	3.5
TC4	Audio	68.66	362.37	4.31%	3.4
	Video	489.16	316.01	2.39%	3.47
TC5	Audio	68.72	352.57	4.33%	3.33
	Video	276.35	379.55	2.38%	3.37
TC6	Audio	68.67	361.58	4.31%	3.4
	Video	524.93	317.29	2.32%	3.53
TC7	Audio	68.57	357.84	5.05%	3.13
	Video	322.84	321.77	3.13%	3.6
TC8	Audio	68.48	359.58	5.17%	3.2
	Video	524.3	306.79	2.55%	3.63

Table 6.7: Correlations between QoE and perceived audio- and video quality.

Pearson correlation	Perceived audio quality	Perceived video quality
QoE	0.779**	0.716**

** Correlation is significant at the 0.01 level

6.1.3 QoE model for network impaired session derived from data collected in user study US3

QoE for multiparty audiovisual telemeetings in general has numerous important influence factors, as previously discussed. Packet loss and delay belong to key factors related to the system used for service delivery. WebRTC video conferencing services have implemented a congestion control mechanism which adapts video quality in response to packet loss and delay. Preset parameters (resolution, bitrate, and frame rate, directly related to the perceived quality), are dynamically adapted in response to measured network conditions. In case of network impairments, the duration and the level of quality degradation as well as the beginning and the ending of the impaired session can impact the overall experience. Thus, the following models may not be conclusive, since the experiment was not designed with the power to evaluate the indirect relation of the packet loss and delay with the video quality parameters (bitrate, resolution, frame rate), but they can be very useful to generate further experiment and guide future studies.

To develop regression models for QoE estimation of the impact of network impairments in terms of audio and video packet loss and delay, we used collected data in user study US3. It should be noted that some of the data sets were not well-modeled by a normal distribution and showed skewness and kurtosis, but the ANOVA test is considered as robust against normality assumptions [149]. Also, dependent variables are measured as interval, since the difference between the points on the rating scale is considered to be equal.

Additionally, linear correlations between perceived video and audio quality and QoE were measured and reported in Table 6.7. Pearson's product moment correlation coefficient r was computed. Results show significant positive correlation between both, QoE and perceived audio quality, as well as QoE and perceived video quality.

Perceived video quality (PVQ), perceived audio quality (PAQ), and overall quality ratings (QoE) reported in user study US3 served as an input to derive quality models in case of network disturbances corresponding to short bursty packet loss and application layer delay. To obtain higher accuracy, we based the model on the 36 unique 3-tuple values (audio-, video-, overall quality ratings) extracted from 240 measurements collected in user study US3 (across all test scenarios) (Figure 6.5). We tried to fit the data using various polynomial, rational, and linear models, however results did not show significantly greater accuracy between models. The QoE

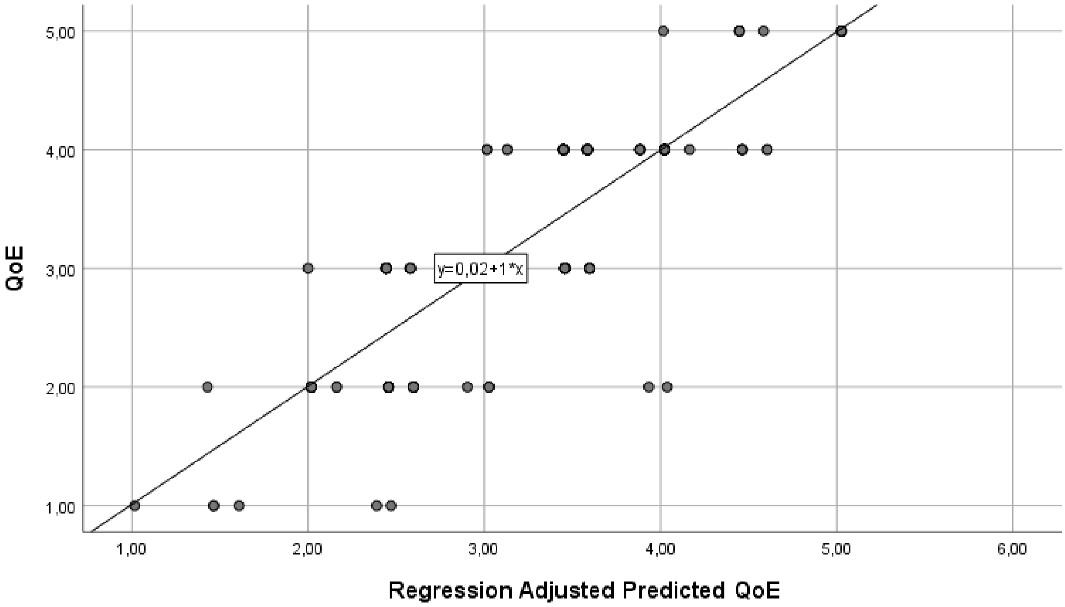


Figure 6.5: Accuracy of estimated overall quality ratings (horizontal axis) compared to the actual overall QoE ratings (vertical axis) collected in user study US3 (based on the model given in Equation 6.1).

model equation based on linear regression with two features PAQ and PVQ as predictors is calculated as follows:

$$QoE = 0.57 * PAQ + 0.433 * PVQ - 0.012 \tag{6.1}$$

Taken as a group, the predictors PVQ and PAQ account for 74.4% of the variance in overall QoE. Both predictors, PAQ and PVQ have a statistically significant impact on the QoE because their p-values are less than the significance level of 0.001.

Based on the generic model (eq. 3.1), we obtained following equation for the QoE$_{\text{gen}}$ model:

$$QoE_{gen} = 0.643 * PAQ + 0.495 * PVQ - 0.02 * PAQ * PVQ - 0.203 \tag{6.2}$$

The coefficient of determination R^2 indicates that 74.6% of the total variance is explained by the independent variables PAQ and PVQ. Adding additional multiplication factor into equation did not contribute significantly to the greater model accuracy.

Quadratic polynomial model yielded an even higher R^2 value of 0.747. However, insignificantly increased accuracy does not justify model complexity.

PAQ was modeled with audio delay and packet loss (eq. 6.3), and the obtained results show that taken as a group, predictors audio delay (ADL) and audio packet loss (APL) account for 89.2% of the variance in perceived audio quality. Both predictors, APL and ADL, have a statistically significant impact on the perceived audio quality because their p-values are less than the significance level of 0.05. The accuracy of the PAQ prediction model is shown in Figure

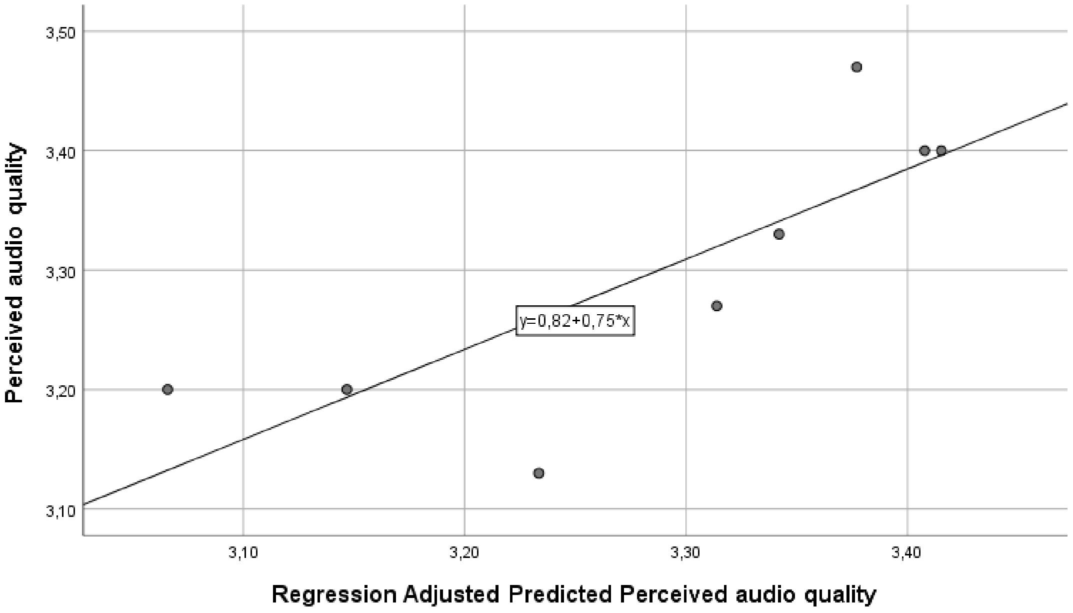


Figure 6.6: Accuracy of predicted audio quality ratings (horizontal axis) compared to the actual audio quality ratings (vertical axis) collected in user study US3 (based on the model given in Equation 6.3).

6.6.

$$PAQ = 0.007 * ADL - 0.253 * APL + 2.046 \tag{6.3}$$

PVQ was also modeled with video delay (VDL) and packet loss (VPL) as predictors. Results showed that taken as a group VDL and VPL account for 80.8% of the variance in perceived video quality. The predictor video delay is statistically significant because its p-value is less than the significance level of 0.05, while the predictor video packet loss is statistically insignificant ($p > 0.05$). The accuracy of the PVQ estimation model is shown in Figure 6.7.

$$PVQ = -0.003 * VDL + 0.106 * VPL + 4.066 \tag{6.4}$$

Summary of key findings

Performance measurements showed that packet loss caused severe disturbances during multiparty audiovisual calls established via smartphone devices, in some cases even the reduction of video bitrate to nearly zero. The impact of a “lost” video stream on overall QoE was found to differ greatly among participants, which can be attributed to differences in end user expectations. As long as the audio quality remained satisfactory, most participants provided relatively high quality scores. Considering that audio was not lost in any sessions, we can conclude that in a leisure conversational context, where participants are also acquaintances, temporary video loss may not present a strong negative impact.

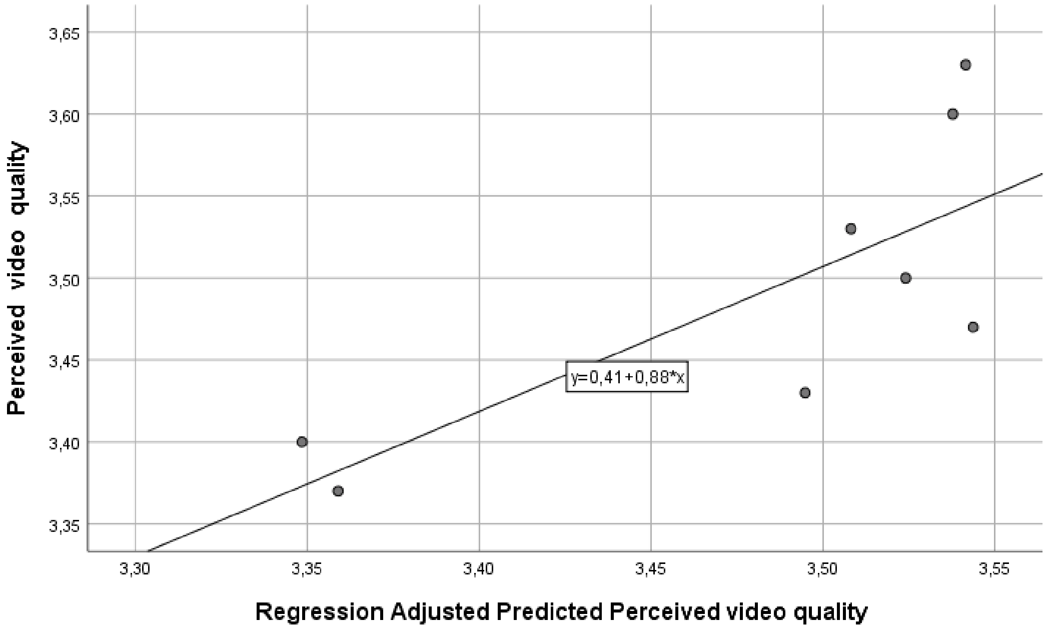


Figure 6.7: Accuracy of estimated video quality ratings (horizontal axis) compared to the actual video quality ratings (vertical axis) collected in user study US3 (based on the model given in Equation 6.4).

6.2 User study US4

The purpose of user study US4 was to use it as a baseline study, so as to compare against data collected in study US3 in order to investigate the difference in perceived quality in cases with (US3) and without (US4) network disturbances (packet loss and application layer delay). Thus, in user study US4 we investigated how the same video encoding parameters (in terms of encoding bitrate, resolution, and frame rate) as in study US3 but without network impairments (packet loss and delay on application layer), influence QoE (results reported in [41]). In user study US4 we tested limited bandwidth scenarios by limiting encoding bitrate. Therefore, this user study helps to answer the high-level research question RQ5 as defined in Figure 1.1.

6.2.1 Methodology

WebRTC video calls, based on UDP/RTP protocols, were established between three participants, located in three different rooms, as illustrated in Figure 6.8. Participants were connected via a 2.4GHz and 802.11n WiFi router ASUS RT-AC51U, while the Licode media server connected via a fixed connection to the router. To analyze the adaptation of encoding parameters and to monitor network conditions, we collected *webrtc-internals* statistics, including the *getUserMedia* information (session id and origin as well as passed audio and video constraints, such as video width, height, frame rate) along with *RTCPeerConnection* data (performance monitor data such as send and receive bitrate, available bandwidth, delay, packet loss, quality limitation reason, etc.) for each participant [147].

In the experiments, video resolution, bitrate and frame rate streamed by each client were predefined using settings in the Licode media server, installed in a local network on a computer with Intel Core i5 Processor, 2.6 GHz, 8 GB RAM and Ubuntu 14.04 LTS. We set-up a local and symmetric environment, i.e., all participants used the same end user device: Samsung Galaxy S6 with 3GB RAM, display size 5.1" and display resolution 1080x1920 px. With respect to traffic flows, each participant had one outgoing audio and one outgoing video flow, and two incoming audio/video flows.

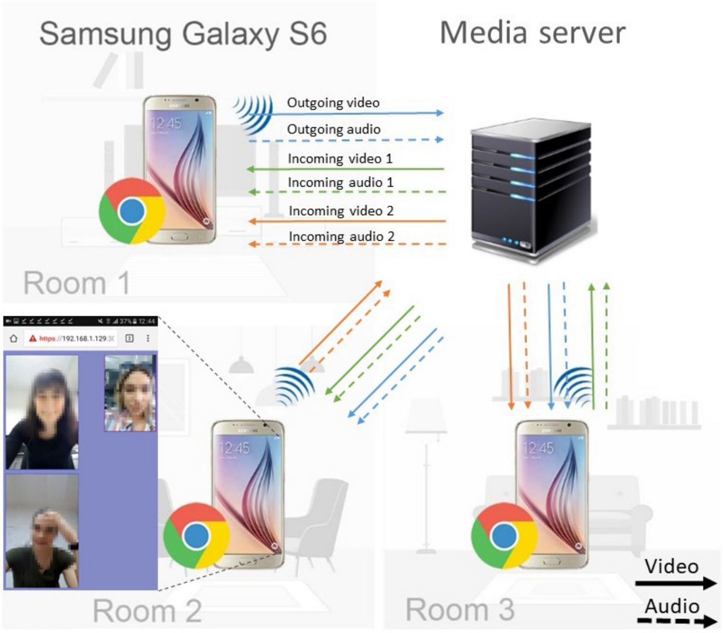


Figure 6.8: Testbed set-up over a LAN (user study US4).

The test schedule consisted of each user group testing 8 conditions based on the following: 320x480 px and 480x640 px video resolutions encoded with VP8 video codec, encoding bitrates set to 300 kbps and 600 kbps, and with frame rate set to 15 fps and 20 fps (Table 6.8). The settings refer to each video stream within a call, and were always set symmetrically, i.e., the same for all client streams. With 8 test conditions and 9 user groups (each group with three participants), this resulted in a total of 72 performed tests.

Established video calls lasted for three minutes per test case and were initiated through a custom made WebRTC application within the Google Chrome 57.0.2987.132 browser. At the beginning of the testing session, a preliminary test was carried out to familiarize participants with the assignment and assessment questionnaire. Preliminary results are not taken into account. After the completion of each condition, subjects were asked to rate overall quality, audio quality, video quality, AV synchronization using a paper questionnaire and the 5-pt. ACR scale. Participants were asked to use their native language and free conversation without any predefined task.

Twenty-seven (21 male and 6 female) participants took part in the study on a voluntary basis,

Table 6.8: Test schedule used in user study US4.

Experiment	Video resolution [px]	Frame rate [fps]	Bitrate [kbps]
Test case 1 (TC1)	320x480	15	300
Test case 2 (TC2)	320x480	15	600
Test case 3 (TC3)	320x480	20	300
Test case 4 (TC4)	320x480	20	600
Test case 5 (TC5)	480x640	15	300
Test case 6 (TC6)	480x640	15	600
Test case 7 (TC7)	480x640	20	300
Test case 8 (TC8)	480x640	20	600

with an average age of 21 years (youngest was 20 and the oldest 29 years old). Participants were divided into nine groups, formed based on acquaintances. Twenty-four participants were students and three participants were employed and have participated previously in subjective assessments. The selected subjects reported having previous experience with applications such as Skype, Viber and WhatsApp. All participants have normal or corrected vision and normal hearing.

6.2.2 Results

Even though video coding parameters corresponding to each test condition were set as fixed default values in the Licode media server, we observed during the test cases that these default values were commonly reduced. In general, quality degradation may be caused due to lack of available send or receive bandwidth, delay, packet loss, and/or limited CPU processing power. In our study, we observed that quality reductions were triggered mainly by available bandwidth and CPU overuse. In this study quality was not degraded by the packet loss, since significant packet loss was not recorded for any test case. Values for test cases TC1 to TC4 ranged from 0.009% to the 0.774% (shown in Table 6.9). For test cases TC5 to TC8, packet loss ranged from 0.012% to 2.419% in TC8. We note however that the vast majority of TC8 packet lost occurred in only one test group.

Available receive bandwidth: Receive bandwidth refers to available bandwidth for each video stream, as estimated by the receiving client. When observing average receive bandwidth measurements per test case, values indicated that there was enough receive bandwidth for each incoming video stream. For test cases where we set the video encoding bitrate for each partic-

Table 6.9: WebRTC internals data: mean values of measured parameters (averaged across all test groups) and Mean Opinion Score per test condition.

Measured parameter / Test condition	TC1	TC2	TC3	TC4	TC5	TC6	TC7	TC8
Packet loss (%)	0.009	0.411	0.774	0.013	0.441	0.012	0.309	2.419
Audio delay (ms)	289.25	318.54	299.92	329.46	302.96	315.18	306.91	320.87
Video delay (ms)	241.54	269.09	251.35	279.07	250.91	263.32	256.91	263.81
Minimum receive bandwidth (kbps)	353.75	743.11	395.03	800.65	387.19	770.93	473.91	801.82
Maximum receive bandwidth (kbps)	1050.2	2142.44	777.53	1363.86	1280.43	1564.15	2596.57	2643.29
Available receive bandwidth (kbps)	548.12	1051.98	531.62	969.91	554.26	1032.38	582.91	1041.29
Minimum send bandwidth (kbps)	226.21	432.04	232.21	537.78	235.82	497.67	244.83	260.32
Maximum send bandwidth (kbps)	540.87	599.79	300.03	596.15	300.00	599.68	299.15	595.77
Available send bandwidth (kbps)	309.19	576.01	293.18	576.07	294.07	585.44	294.74	551.74
Actual encoding bitrate (kbps)	304.47	568.04	305.72	586.88	301.56	574.68	298.95	575.12
Percentage of session time where actual streamed resolution corresponded to the set resolution 320x480 px (%)	87.42	95.57	66.64	94.39	-	-	-	-
Percentage of session time where actual streamed resolution corresponded to the decreased 360x240 px resolution (%)	12.58	4.43	33.36	5.61	-	-	-	-
Percentage of session time where actual streamed resolution corresponded to the set resolution 480x640 px (%)	-	-	-	-	6.68	14.22	9.79	10.82
Percentage of session time where actual streamed resolution corresponded to the decreased 480x360 px resolution (%)	-	-	-	-	72.62	85.35	53.38	80.52
Percentage of session time where actual streamed resolution corresponded to the decreased 320x240 px resolution (%)	-	-	-	-	20.70	0.43	36.83	8.66
Percentage of session time where actual streamed resolution corresponded to the set frame rate and +/-1 (%)	96.51	96.03	92.18	98.15	97.61	98.88	97.28	94.04
Mean Opinion Score								
MOS audio quality	3.29	3.48	3.26	3.26	3.07	3.26	3.04	3.70
MOS video quality	3.77	4.18	3.41	4.11	2.70	3.59	2.63	3.93
MOS AV synchronization	4.03	4.07	3.85	3.89	3.74	3.89	3.41	4.04
MOS overall quality	3.62	3.92	3.51	3.85	3.19	3.70	3.15	3.89

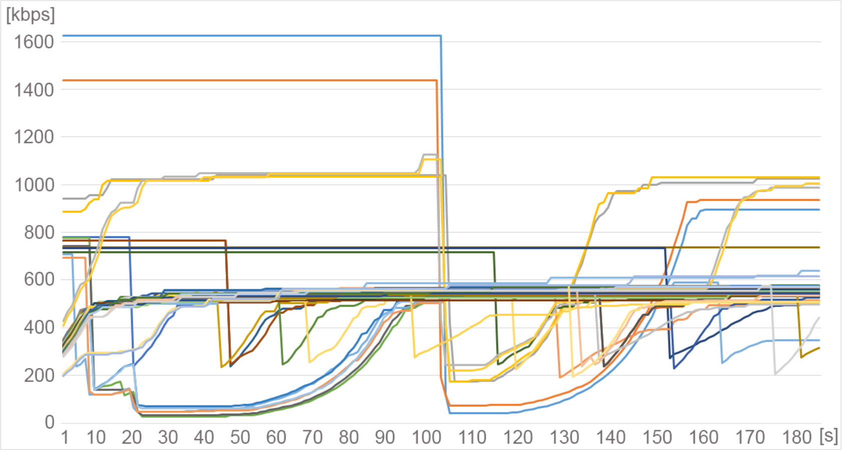


Figure 6.9: TC1 available receive bandwidth per incoming video stream. Each curve represents one of the 27 participants, each receiving two videos. Thus, 54 curves are portrayed.

ipant to 300 kbps (TC1, TC3, TC5, TC7), the average available receive bandwidth was 554.22 kbps, with a minimum average value 353.75 kbps in TC1, shown in Table 6.9. Test cases where encoding bitrate was set to 600 kbps (TC2, TC4, TC6, TC8) also appear to have had enough receive bandwidth for all test cases and groups, with an average of 1023.89 kbps and minimum 743.11 kbps in TC2. However, if we consider per second level receive bandwidth values in every test condition, we observe fluctuations which might have an impact on perceived quality. As an example, we illustrate TC1 measurements which show available receive bandwidth as reported in webrtc-internals for incoming video streams (Fig. 6.9).

Current Delay: Current delay includes jitter buffer, decode time and a render delay, in case of a video stream. Average audio delay values ranged from 289.25 ms (one way) in TC1 to 329.46 ms in TC4. Even though the average delay range was acceptable, deviations occurred in each test case. This occurred despite the fact that all measurements were conducted in an isolated local network, as previously described.

Similar observations resulted in the case of reported video delay, where average values ranged from 241.54 ms in TC1 to 279.07 ms in TC4. For all test cases, video delay was lower on average by 16.51% than audio delay. Video delays, caused by network and video processing, ranged from short and small delays, to long peaks, clearly impacting perceived quality.

Resolution: While in each test case we set the default video resolution values, measurements showed that these values were not maintained during the course of each video call, thus resulting in numerous resolution fluctuations. The default resolution which was maintained for the maximum amount of time (95.57% of the session) was 320x480 px, in TC2. The default resolution maintained for the minimum amount of time was 480x640 px (6.68% of the session) in TC5. Subjects rated TC7 as the worst condition, with lowest encoded resolution 320x240 px held for the longest time, 36.83% of the session. Frame rate default value was stable within all test cases, maintained for the longest time in TC6 with 98.88% of the session and 15 fps, and shortest period in TC3 with 92.18% of the session and 20 fps. Performance analysis showed that increased frame rate did not result in the highest video quality MOS rating. The highest MOS was observed for a frame rate of 15 fps, as shown in Table 6.9.

Resolution limitation: In this study, limitations causing the change of default encoding settings were *send bandwidth* and *CPU*. To obtain more information as to why resolution was lowered, we analyzed the webrtc-internals parameters *send-googBandwidthLimitedResolution* and *sendgoogCpuLimitedResolution*, referring to insufficient bandwidth for stream transmission and CPU overuse. Insufficient send bandwidth seems to have presented a problem for all test cases. Due to the bandwidth limited resolution, the default resolution was degraded within 127 video streams. The most frequent occurrence was observed in TC5 (25 out of 27 streams were degraded with respect to default resolution that was set), and in TC7 (26 out of 27 streams were degraded with respect to default resolution that was set) (Table 6.10).

Table 6.10: WebRTC internals data of average, minimum and maximum values per test condition, occurrence per sent stream.

Test case	Bandwidth Limited Resolution		CPU Limited Resolution	
	Occurred	Avg/Min/Max duration (s)	Occurred	Avg/Min/Max duration (s)
TC1	9/27	61.87 / 12 / 180	0/27	0 / 0 / 0
TC2	8/27	32.87 / 4 / 83	0/27	0 / 0 / 0
TC3	14/27	131.14 / 24 / 179	0/27	0 / 0 / 0
TC4	10/27	34.4 / 4 / 135	0/27	0 / 0 / 0
TC5	25/27	172.16 / 20 / 179	2/27	90.5 / 42 / 139
TC6	18/27	88.72 / 12/ 179	20/27	137.5 / 56 / 179
TC7	26/27	166.88 / 14 / 179	5/27	81.8 / 6 / 157
TC8	18/27	102.27 / 8 / 179	20/27	147.1 / 35 / 179

In all test cases, video resolution was scaled down due to unavailable send bandwidth or motion factor. In general, for test cases requiring 300 kbps, resolution lowering was caused more often because of a motion factor, especially for 480x640 px resolution. Results showed that with an average 554.27 kbps available receive bandwidth per participant, the designated 300 kbps for 480x640 px resolution and 20 fps is a borderline value, since in two groups 480x640 px resolution was maintained for 113 and 105 seconds. Regardless of the fact that a video call should be a low motion service, for other groups, 300 kbps was not enough. On the other hand, test cases requiring 600 kbps lowered resolution more frequently due to reported lack of send bandwidth.

CPU limited resolution was not triggered within test cases with the lower default resolution

320x480 px, only with 480x640 px, in particular within test cases with predefined 600 kbps encoding bitrate. In TC6 and TC8, the video resolution of sent videos was lowered in 20 out of 27 participants for both test cases. In sessions lasting 180 seconds, TC6 resolution was on average lowered for a duration of 137.5 seconds, while in TC7 resolution degradation lasted for 147.1 seconds, shown in Table 6.10. In test cases with 300 kbps, 2 out of 27 participants experienced a CPU limitation in TC5, where video being sent on average was encoded with lowered resolution for a duration of 90.5 seconds. In TC7, resolution was lowered for 81.8 seconds on average within 5 out of 27 participants.

Our measurements show that 480x640 px resolution has a strong impact on CPU utilization in cases when utilizing high-end smartphones (with 3GB RAM) in a three-party video call, especially in case of higher video bitrates, and should thus be avoided.

QoE metrics

To obtain better insights into subjective ratings, rating distributions were calculated per test condition and are shown in Fig. 6.10.

In TC1 audio quality was rated with MOS value 3.29 and video quality 3.77, where 3.7% of unsatisfied participants rated audio and video quality with “Bad”, while 7.4% and 18.52% participants rated audio and video as “Excellent”, respectively. The lowest synchronization rating was “Fair” reported by 18.52% of participants, while “Excellent” was reported by 22.22% of the participants. Overall quality was most often rated as “Fair”, with a share of 44.44%.

For TC2, calculated MOS for audio quality scored 3.48, video quality was rated with MOS 4.18 and thus slightly higher than TC1, synchronization 4.07 and overall quality 3.92. Further comparing TC3 to TC1 with difference being in higher frame rate, TC3 had slightly lower ratings per category but was never rated as “Bad”. MOS values for TC3 correspond to 3.26 for audio quality, 3.41 for video quality, 3.85 for synchronization and 3.51 for overall quality. The most frequent rating for audio, video and overall quality was “Fair”, while synchronization was rated as “Good” by 48.15% of participants.

Comparing TC4 to TC2 with a 5 fps difference, as in the case with TC3 and TC1, TC4 scored slightly lower ratings, 3.26 MOS for audio, 4.11 for video, and 3.85 for overall quality along with 3.89 for synchronization. In this test case, unlike TC1, participants never rated the test condition as “Bad”. For video quality, the most common rating was “Excellent” given by 40.74% of participants. However, the most frequent rating for audio quality was “Poor” with a share of 33.33% participants.

TC5 and TC7 were cases with higher resolution and lower video bitrate. Both test cases scored the lowest ratings. TC7 yielded the lowest MOS values for overall quality (3.15), audio (3.04) and video (2.63) quality as well as synchronization (3.41). Participants reported in TC5 “Bad” quality for audio and video, while in TC7 participants perceived synchronization as “Bad” as well.

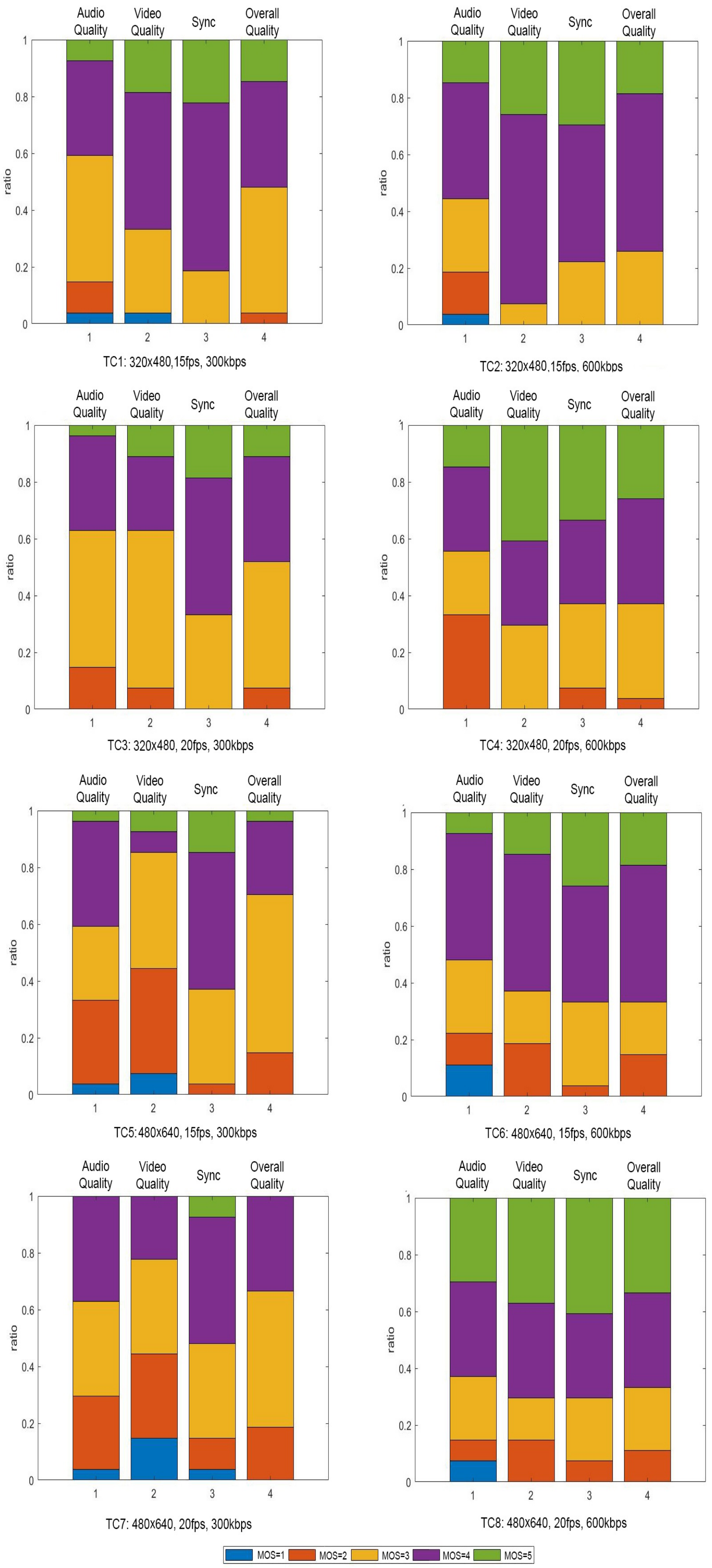


Figure 6.10: MOS distributions for audio quality, video quality, AV synchronization and overall quality per test condition.

Conditions measured in TC6 and TC8 both have occurrences of “Bad” audio quality, in 11.11% and 7.4% of cases (respectively), with MOS values of 3.26 (TC6) and 3.70 (TC8). All rated categories yielded certain percentages of “Excellent” scores, occurring in TC6 in case of audio quality in 7.4%, video quality 14.81%, synchronization 25.93% and overall quality 18.52%. In TC8, the rating “Excellent” was given in case of audio quality 29.63%, video quality 37.04%, overall quality 33.33%, and 40.74% for synchronization.

Significant differences between ratings caused by unexpected disturbances showed that video encoding adaptation can be proactive or reactive. Proactive implies environment scanning in terms of constant parameters, such as end user device capabilities, while reactive implies in response to network impairments. We find that test cases with predefined default resolution 320x480 px gained higher scores given by both unsatisfied and satisfied participants implying that such test conditions are manageable for tested smartphones. Two test cases evaluated with the highest average scores, TC2 and TC4, suggest that 480x640 px resolution is too high and 300 kbps too low for three-party audiovisual telemeeting in case of the delay, bandwidth and CPU limitations.

Sessions with lowest MOS ratings per test condition

Since our test cases did not result in stable video encoding settings maintained for the duration of each session, as we had expected given our controlled lab environment, we further specifically analyze the session conditions of two incoming video and audio streams from the perspective of the most unsatisfied participant (i.e., lowest ratings) so as to identify what might have caused user annoyance.

In TC1, for both incoming streams, resolution was not lowered from the beginning to the end of the session. The level of available *receive bandwidth* was in case of one incoming stream more than enough for the whole session duration, while in the other case *receive bandwidth* was sufficient and stable until the 160th second, when audio and video delay started to increase along with the bandwidth reduction which caused video bitrate reduction. As a consequence, video quality was rated as a “Bad”, overall quality “Poor”, audio quality “Fair” and AV synchronization “Good”.

TC2 did not experience default resolution changes as well. Several reductions of available *receive bandwidth* occurred for both streams, with one longer and under 600 kbps reduction inducing video bitrate lowering. Added higher audio and video delay resulted with “Bad” audio, “Fair” AV synchronization and overall quality and surprisingly high “Good” video quality.

With the sufficient available *receive bandwidth*, TC3 encountered resolution reduction in both incoming streams due to the send bandwidth, with one being reduced to resolution 360x240 px and lasting 87.77% of the session duration time. Audio and video received bitrate was acceptable for both streams, with one short video bitrate drop, shaping participant’s “Poor” opinion for audio, video and overall quality and “Fair” for AV synchronization.

TC4 managed to maintain the default resolution with sufficient received bandwidth (including two shorter drops) and acceptable incoming bitrates, but approximately 400 ms audio and video delay led to the “Poor” QoE rating of audio and overall quality and “Fair” rating of video quality and AV synchronization.

TC5 had enough received bandwidth and acceptable both bitrates, but the predefined resolution was decreased to 480x360 px and retained during the whole session. Since the available send bandwidth was as predefined 300 kbps, resolution was decreased due to the inability to encode 480x640 px resolution within 300 kbps. Considering significant audio and video delay as well, low ratings were given corresponding to “Bad” perceived audio and video quality, “Poor” overall quality, and “Fair” AV synchronization.

In TC6, available receive bandwidth was sufficient for all incoming streams with two short and not so severe drops for one participant’s streams, with fairly stable incoming bitrates. Predefined resolution was maintained at one incoming stream for 42.77% of session time, while the other incoming stream had a lowered 480x360 px resolution during the whole session. *Send bandwidth limitation* occurred due to the encoding issues, causing the whole session resolution reduction, while the CPU limitation caused partial reduction.

Available receive bandwidth in TC7 was sufficient for incoming streams (at least 500 kbps), except one reduction occurrence (for only one participant’s incoming stream) to the less than 300 kbps lasting for approximately 10 seconds. Bitrate was stable for streams from both participants carrying the default resolution until the end. High and long-lasting video and audio delay was present especially for streams coming from one participant, causing freezes in the communication several times. The most unsatisfied participant rated the session AV synchronization as “Bad” and audio, video, and overall quality as “Poor” .

TC8 had sufficient receive bandwidth with one short disturbance per incoming stream from each participant, with acceptable audio and video bitrates. For one incoming stream, the resolution was lowered to 480x360 px for 69.44% of the session time. For the other incoming stream, the resolution was degraded for 87.22% of the session. Resolution was degraded for the first incoming video stream only because of CPU limitation, and in case of the second incoming stream resolution was firstly reduced because of the predefined send bandwidth limitation, and after short recovery, again decreased due to the CPU limitations. Significantly high video and audio delay occurred in all incoming streams, with an appearance at the end of the session in one case which could result in a memory effect and impact user judgment. Such quality degradation produced “Bad” audio quality and “Poor” AV synchronization, video, and overall quality.

Summarized results showed that test cases where we set default lower video resolutions to 320x480 px, degradation was not triggered due to the CPU limitation, as was the case with test conditions based on the 480x640 px resolution. Resolution was automatically lowered because of a lack of send bandwidth, which occurred more often with test conditions requiring 600 kbps, while 300 kbps showed sensitivity to the content motion.

Summary of key findings

Based on our conducted tests and analysis of obtained results in US4, we draw the following conclusions:

- For a multiparty (3-way) video call established via mobile devices, 480x640 px resolution per participant window is not recommended for smartphones with 3GB or less, given the CPU usage requirements.
- If bandwidth is limited, both resolution and video bitrate should be reduced. Based on the results of our studies, we conclude that if bandwidth is below 400 kbps, 300 kbps or 200 kbps, encoding bitrate and resolution should be set to 350 kbps and 320x480 px, 250 kbps and 240x360 px, 150 kbps and 180x240 px respectively, so as to avoid quality degradation due to delay and loss occurrence. Even though the GCC algorithm is designed to mitigate such quality degradations by adapting the bitrate, resolution, and frame rate, our goal is to identify those codec settings that result in satisfactory QoE, while at the same time proactively attempting to avoid situations that would results in the need to trigger the GCC algorithm.

6.3 Chapter summary

The goal of conducted studies has been to investigate the impact of a network impaired environment (involving packet loss and application layer delay in US3) and a network unimpaired environment (providing baseline scores reported in US4) on the perceived quality, by comparing scores obtained in different test scenarios in terms of video codec configuration settings. We note that a different group of participants took part in study US3 as compared to US4, so we refrain from drawing concrete conclusions regarding quantification of differences in ratings across test conditions. However, by comparing ratings we are able to obtain initial insights into packet loss-QoE relationship. Results showed that no significant differences in subjective ratings exist between the same test conditions with and without packet loss. Further data analysis indicates that quality reduction caused by temporarily inserted packet loss and activation of GCC algorithm is lower and lasts for a shorter time period in cases when sessions were configured with lower default video quality (in terms of resolution, fps, bitrate) as compared to sessions originally configured to stream higher video quality. In certain cases, adaptation led to video interruption. In majority of other cases, we observed that it took approximately 25 seconds for the video stream to recover to an acceptable quality level after the temporary occurrence of network packet loss. While video loss had a significant impact on certain participants, for majority of participants it did not significantly contribute to the perceived quality. In the following chapter, we described video encoding adaptation strategies for multiparty video calls on mobile devices based on the findings from conducted user studies and on proposed quality estimation models.

Chapter 7

QoE-aware video quality adaptation

In this chapter, we first describe how to determine video encoding parameters (bitrate, resolution, frame rate) in accordance with mobile device processing capabilities (section 7.1). Subsequently, these results are utilized for the purpose of deriving video encoding adaptation strategies presented in section 7.2 in the second part of the chapter.

7.1 Recommendation on video encoding parameters for multiparty calls on mobile devices

7.1.1 Smartphone capabilities versus video encoding parameters

User ratings are impacted by typical motion in context of a video call on mobile devices. Participants taking part in our studies were encouraged to act normally as they would in real life scenario. They were allowed to hold the smartphone in their hand or leave it on a stand, as we wanted to capture motion levels specific for audiovisual telemeetings.

We analyzed two parameters from *webrtc-internals*, *send-googBandwidthLimitedResolution* and *send-googCpuLimitedResolution* which provide information on whether the resolution was adapted due to CPU issues, or if there was not enough available bandwidth (in this case referring to encoding bitrate). Both parameters will return true if adaptation occurs, otherwise the parameter value will be false. We selected test cases where resolution adaptation, based on CPU overuse and insufficient bandwidth, occurred in less than 5% of all sessions per particular test scenario (Table 7.1). We used 328 logs for analysis, since we had to discard some logs due to incompleteness and errors. In the majority of selected cases packet loss was zero or very small, less than 2% on average. Average encode time of video frame ranged from 9.46 ms, for the test case with preset resolution 120x180 px, 15 fps and 100 kbps, to 31.34 ms for the test case with preset resolution 320x430 px, 20 fps and 300 kbps. Encoding usage presents encode time per frame divided by the average collection time per frame, and it ranged from 23.55 - 62.28%.

Table 7.1: Selected test cases where resolution adaptation occurred in less than 5% of all sessions.

Average encode time [ms]	Average encode usage [%]	Average actual encoding bitrate [bps]	Preset bitrate [bps]	Resolution [px]	Frame rate [fps]	Resolution x Frame rate
9.46	23.55	97238.77	100000	120x180	15	324000
11.42	28.37	196634.25	200000	120x180	20	432000
19.96	46.18	294688.72	300000	180x240	10	432000
15.05	37.34	198126.30	200000	180x240	15	648000
15.86	38.49	294698.55	300000	180x240	20	864000
22.06	58.67	147540.50	150000	240x360	15	1296000
23.17	61.34	197220.55	200000	240x360	15	1296000
26.38	58.95	294711.73	300000	240x320	20	1536000
21.67	57.57	289496.05	300000	240x360	20	1728000
28.74	63.36	599543.17	600000	320x480	15	2304000
31.34	68.28	292943.77	300000	320x430	20	2752000
28.21	61.51	603063.79	600000	320x480	20	3072000

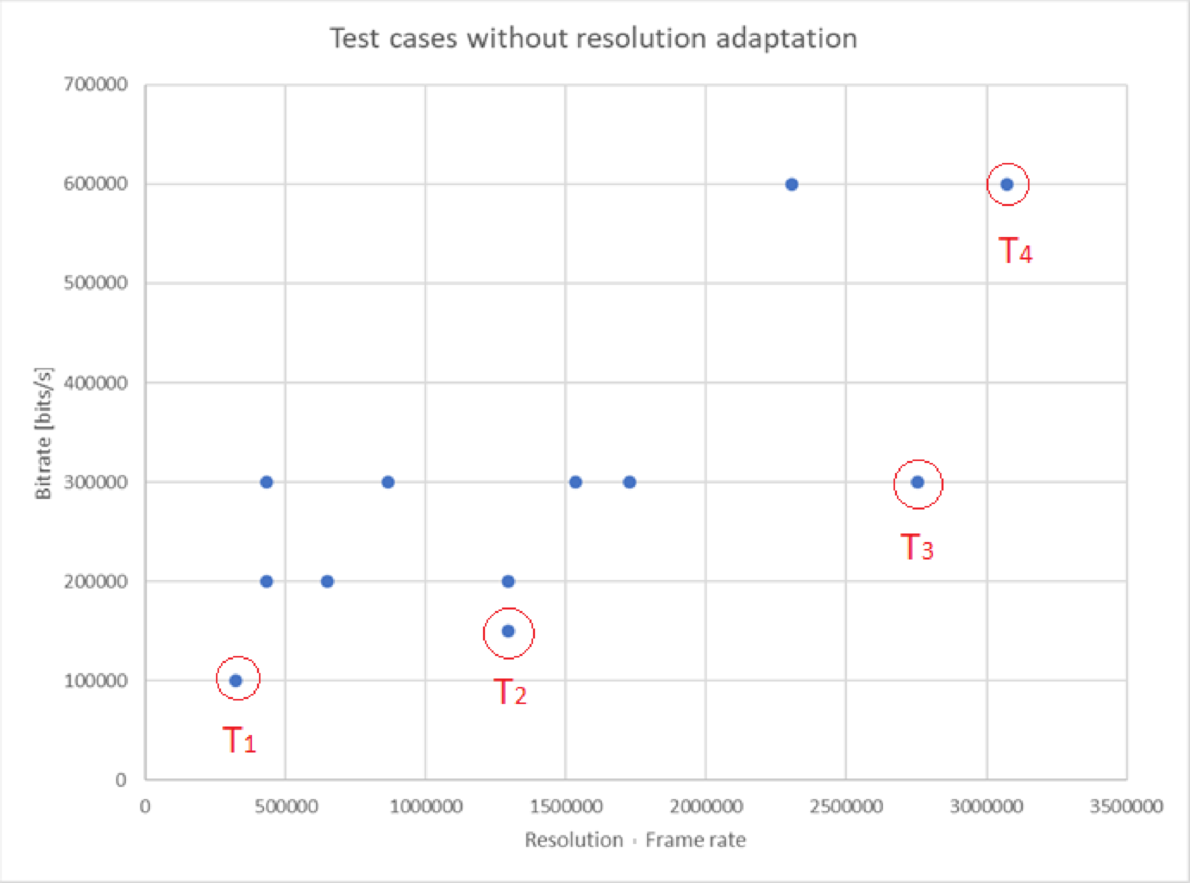


Figure 7.1: Selected test cases where resolution adaptation occurred in less than 5% of all sessions per test condition.

We consider *Resolution (Frame height · Frame width) · Frame rate* multiplication as an independent variable, plotted on the X-axis, while preset bitrate is plotted on the Y-axis (Figure 7.1). To identify the highest resolution which can be encoded with a particular bitrate without being adapted during the video call, we selected 4 points from Figure 7.1 with highest *Resolution · Frame rate* multiplication value and lowest bitrate, namely $T_1(324000, 100000)$, $T_2(1296000, 150000)$, $T_3(2752000, 300000)$ and $T_4(3072000, 600000)$. We tried to form a trend line with classical interpolation, exponential and polynomial function with order 2, but we got the inflection and generally functions did not fit well (Figure 7.2, eq. 7.1, 7.2).

$$y = 87516 \cdot e^{5 \cdot 10^{-7} \cdot x} \tag{7.1}$$

$$y = 8 \cdot 10^{-8} \cdot x^2 - 0.1079 \cdot x + 116671 \tag{7.2}$$

Since exponential and quadratic interpolation did not approximate the function well, we used interpolation with a rational quadratic function [154]. We had to eliminate one point, as it would be inconvenient to find a functional dependency in the form *y=f(x)* due to the third order

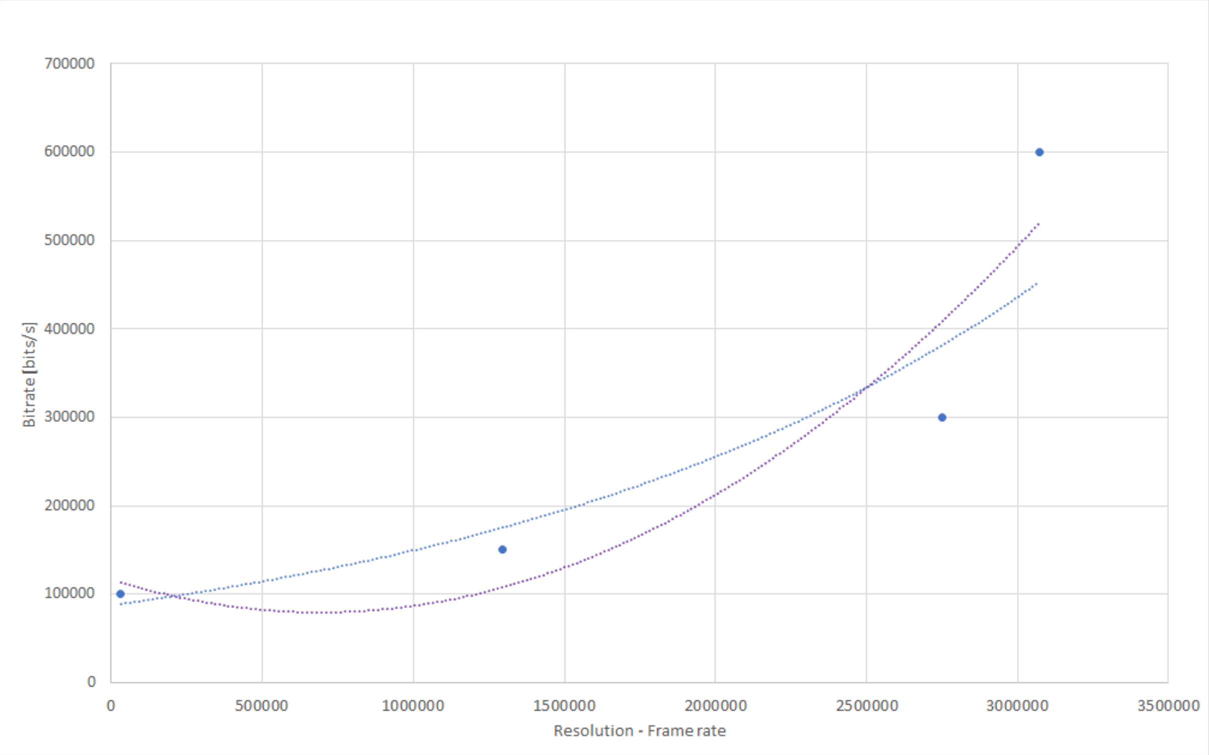


Figure 7.2: Exponential (blue line) and polynomial (purple line) interpolation in 4 points.

parameter t in the parametric form of the equations. Considering point (324000,100000) as less important (participants rated video quality with a mean score 2.35) compared to the remaining three points, we made an interpolation with a quadratic rational function through the points T_2(1296000,150000), T_3(2752000,300000) and T_4(3072000,600000).

Homogeneous coordinates and homogeneous space

The homogeneous space of an n-dimensional space is an n+1-dimensional space. We will strictly limit our workspace to number of dimension *n=2* and the corresponding three-dimensional homogeneous space. Let the coordinates in the workspace of the point *T* be *(x, y)*. The mapping *(x,y) -> (x',y',x_h)* associates to the point *T* an infinite number of points in a homogeneous space according to the following rules:

$$x \cdot x_h = x' \tag{7.3}$$

$$y \cdot x_h = y' \tag{7.4}$$

where the value of x_h is chosen arbitrarily (due to the calculation simplicity x_h=*1* is most convenient to choose).

Interpolation with a rational quadratic function

In cases where we have a curve whose analytic expression is not known, but instead has several points defined, the procedure of curve interpolation can be applied. Interpolation is used to construct additional data points within range of the discrete set of already defined points [155]. This procedure can be used in various technical fields, and it is typically used in computer graphics. We apply this procedure in our work, as outlined in [155].

The selected three points are the points of the rational quadratic function with the following parametric equations:

$$x = \frac{a_1 \cdot t^2 + b_1 \cdot t + c_1}{a^2 + b + c} \tag{7.5}$$

$$y = \frac{a_2 \cdot t^2 + b_2 \cdot t + c_2}{a^2 + b + c} \tag{7.6}$$

Let

$$x_h = a^2 + b + c \tag{7.7}$$

$$x' = a_1 \cdot t^2 + b_1 \cdot t + c_1 \tag{7.8}$$

$$y' = a_2 \cdot t^2 + b_2 \cdot t + c_2. \tag{7.9}$$

The above expressions correspond to the transformation from a workspace to a homogeneous space. Also, the convenience of matrix representation can be recognized in the following form:

$$(x', y', x_h) = \begin{bmatrix} t^2 & t & 1 \end{bmatrix} * K \tag{7.10}$$

Where the matrix K is a characteristic curve matrix and it has to be determined.

$$K = \begin{bmatrix} a_1 & a_2 & a \\ b_1 & b_2 & b \\ c_1 & c_2 & c \end{bmatrix} \tag{7.11}$$

Since K is a third order square matrix we need three points with the parameters t_1, t_2 and t_3 to obtain a matrix equation according to expression 7.10:

$$\begin{bmatrix} x'_1, y'_1, x_h \\ x'_2, y'_2, x_h \\ x'_3, y'_3, x_h \end{bmatrix} = \begin{bmatrix} t_1^2 & t_1 & 1 \\ t_2^2 & t_2 & 1 \\ t_3^2 & t_3 & 1 \end{bmatrix} * K \tag{7.12}$$

If we select x_h=*1* [1], according to expressions for the transition from homogeneous coordinates to the workspace, equation 7.12 takes the form:

$$\begin{bmatrix} x_1, y_1, 1 \\ x_2, y_2, 1 \\ x_3, y_3, 1 \end{bmatrix} = \begin{bmatrix} t_1^2 & t_1 & 1 \\ t_2^2 & t_2 & 1 \\ t_3^2 & t_3 & 1 \end{bmatrix} * K \tag{7.13}$$

Let the point with the lowest *X* coordinate have the parameter t_1*=0*, and the point with the highest *X* coordinate t_3*=1*. The middle point then has the parameter t_2. The parameter t_2 corresponds to the ratio of the interpolation curve arc length from the point with parameter 0 to that point, and the arc length from the point with parameter 0 to the point with parameter 1. As the interpolation curve is not known, the parameter t_2 should be varied to obtain a satisfactory interpolation curve.

However, one way to give a rough estimation of the parameter t_2 is to draw the lines between adjacent points (instead of the interpolation curve) to determine the value of the parameter according to the procedure described above. In our case this method approximates t_2*=0.6*. This value of the parameter t_2 is the first that has to be examined. After these considerations 7.13 takes the form:

$$\begin{bmatrix} x_1, y_1, 1 \\ x_2, y_2, 1 \\ x_3, y_3, 1 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 \\ t_2^2 & t_2 & 1 \\ 1 & 1 & 1 \end{bmatrix} * K \tag{7.14}$$

From equation 7.14 it is possible to obtain an explicit expression for the unknown matrix *K* in the following way:

$$K = \begin{bmatrix} 0 & 0 & 1 \\ t_2^2 & t_2 & 1 \\ 1 & 1 & 1 \end{bmatrix}^{-1} \begin{bmatrix} x_1, y_1, 1 \\ x_2, y_2, 1 \\ x_3, y_3, 1 \end{bmatrix} \tag{7.15}$$

Expression 7.15 is an explicit expression for the characteristic matrix of the curve which has to be inverted to finally determine matrix *K*. By inserting the coordinates of the points and

[1] This can be achieved if *a=b=0* and *c=1*, which will be a fundamental accuracy check of the obtained matrix *K*.

t_2=0.6 we obtain:

$$K = \begin{bmatrix} 0 & 0 & 1 \\ 0.36 & 0.6 & 1 \\ 1 & 1 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 1296000 & 150000 & 1 \\ 2752000 & 300000 & 1 \\ 3072000 & 600000 & 1 \end{bmatrix} \tag{7.16}$$

$$\Leftrightarrow K = \begin{bmatrix} -1626666.66 & 500000 & 0 \\ 3402666.66 & -50000 & 0 \\ 1296000 & 50000 & 1 \end{bmatrix} \tag{7.17}$$

From expression 7.17 it can be seen that both conditions *a=b=0* and *c=1* are met, so the parametric equations of the interpolation curve in the workspace can be written in the form of expression 7.10, taking into account *x = x'* and *y = y'*:

$$\left.\begin{aligned} x &= -1626666.66 \cdot t^2 + 3402666.66 \cdot t + 1296000 \\ y &= 500000 \cdot t^2 - 50000 \cdot t + 150000 \end{aligned}\right\} \tag{7.18}$$

Substituting *t=0.05k* values into the equation, where *k* is a non-negative integer less than or equal to 20, we obtain a series of points in the coordinate plane and parabola which passes through these points (Table 7.2). With this points the interpolation curve is graphically determined (Figure 7.3). With this curve, we got the maximum *Resolution · Frame rate* multiplication that can be transmitted at a certain bitrate without a resolution reduction (due to lack of smartphone processing capabilities or insufficient bandwidth) for smartphones with 3GB or higher in the context of three-party audiovisual telemeeting.

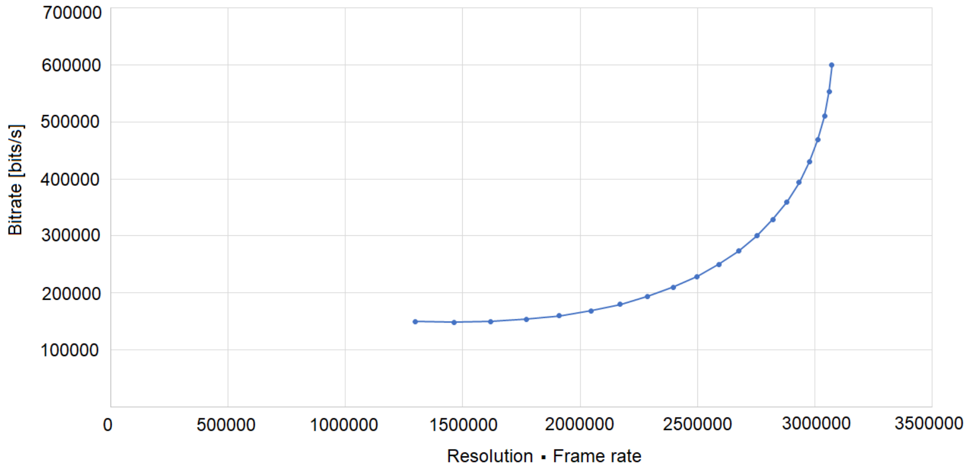


Figure 7.3: Interpolated curve plot for video encoding parameters.

Table 7.2: Interpolated curve points.

t	x	y
0	1296000	150000
0.05	1462066.666	148750
0.1	1619999.999	150000
0.15	1769799.999	153750
0.2	1911466.666	160000
0.25	2044999.999	168750
0.3	2170399.999	180000
0.35	2287666.665	193750
0.4	2396799.998	210000
0.45	2497799.998	228750
0.5	2590666.665	250000
0.55	2675399.998	273750
0.6	2752000	300000
0.65	2820466.665	328750
0.7	2880799.999	360000
0.75	2932999.999	393750
0.8	2977066.666	430000
0.85	3012999.999	468750
0.9	3040799.999	510000
0.95	3060466.666	553750
1	3072000	600000

To obtain the functional dependency in the form *y=f(x)* from the expressions 7.18 it is necessary to eliminate the parameter *t*:

$$y = a_1 \cdot t^2 + a_2 \cdot t + a_3 \tag{7.19}$$

$$x = b_1 \cdot t^2 + b_2 \cdot t + b_3 \tag{7.20}$$

$$b_1 \cdot t^2 + b_2 \cdot t + b_3 - x = 0 \tag{7.21}$$

$$t = \frac{-b_2 \pm \sqrt[2]{b_2^2 + 4 \cdot b_1 \cdot (x - b_3)}}{2 \cdot b_1} \tag{7.22}$$

$$y = a_1 \cdot \frac{x - b_2 \cdot t - b_3}{b_1} + a_2 + a_3 \tag{7.23}$$

$$y = f(x) = a_1 \cdot \frac{x - \frac{-b_2 + \sqrt[2]{b_2^2 + 4 \cdot b_1 \cdot (x - b_3)}}{2 \cdot b_1} b_2 - b_3}{b_1} + a_2 \cdot \frac{-b_2 + \sqrt[2]{b_2^2 + 4 \cdot b_1 \cdot (x - b_3)}}{2 \cdot b_1} + a_3 \tag{7.24}$$

7.1.2 Perceived video quality versus video encoding parameters

The number of subjective ratings of video quality for all tested conditions within user studies US4, US5, and US6 are shown in Figure 7.4, 7.5, and 7.6. Notably lower ratings were reported for scenarios with the lowest tested resolution 120x180 px, and for the highest resolution 480x640 px but evidently preset with insufficient 300 kbps bitrate for the video call on mobile devices. Ratings did not decrease only because of lower resolution, yet due to an insufficient preset bitrate for a particular resolution. It can be noted that the test results show that the impact of frame rate was not significant.

In terms of perceived video quality, taking ratings only into consideration we calculated the percentage of acceptable video quality. Acceptable video quality includes ratings 4 ***"Good"*** and 5 ***"Excellent"***, while unacceptable ratings are 1 ***"Bad"***, 2 ***"Poor"***, and 3 ***"Fair"*** (Table 7.3). Highest rated video quality was in test case 320x480 px, 15 fps, 600 kbps, while the test case 120x180 px, 15 fps, 100 kbps scored lowest video ratings. In a multiparty mobile context, it is not only about bitrate or resolution, and the higher the better. A key challenge is to find the optimal resolution to bitrate ratio, which depends on processing capabilities of each display device, and movement of the camera and/or participant. Thus, to establish the relation between encoding parameters and processing capabilities we have to observe ratings in light of unchanged preset parameters. Highlighted rows mark test scenarios where resolution adaptation

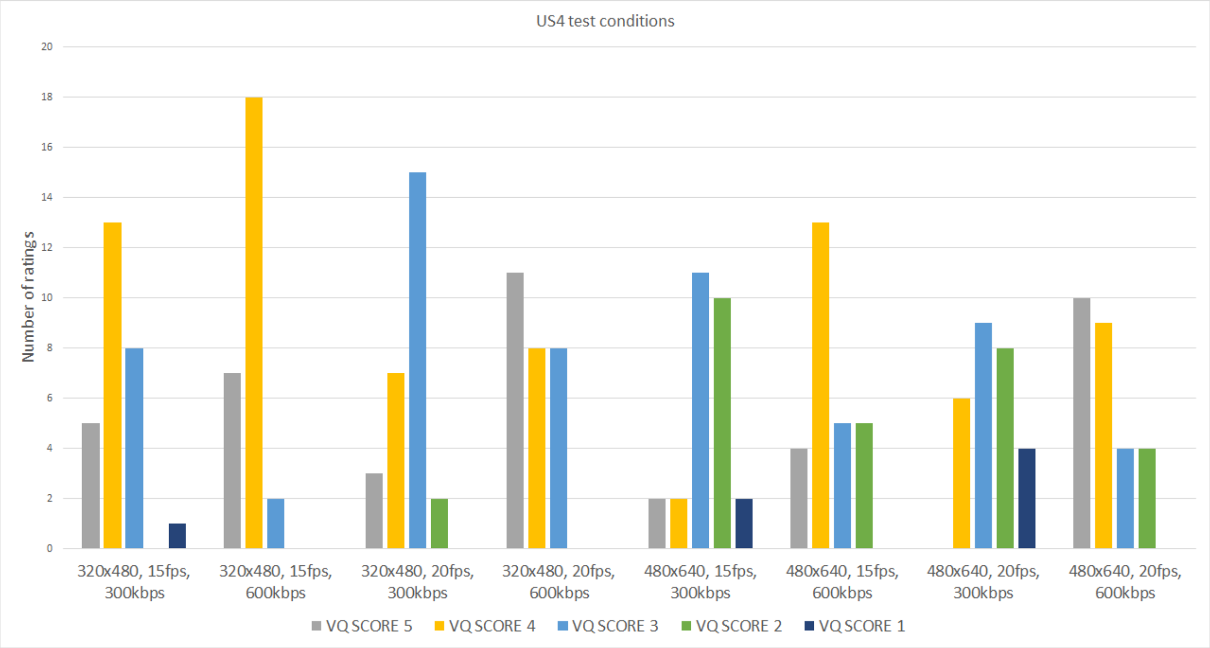


Figure 7.4: Number of reported subjective ratings of video quality for tested conditions in user study US4.

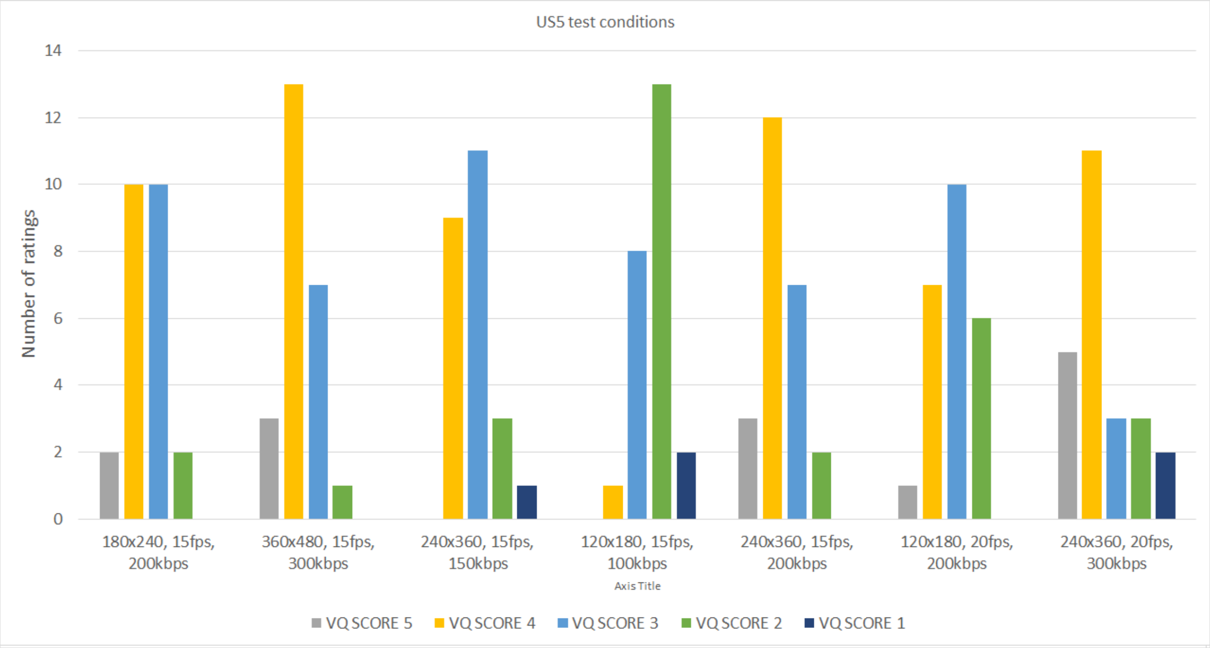


Figure 7.5: Number of reported subjective ratings of video quality for tested conditions in user study US5.

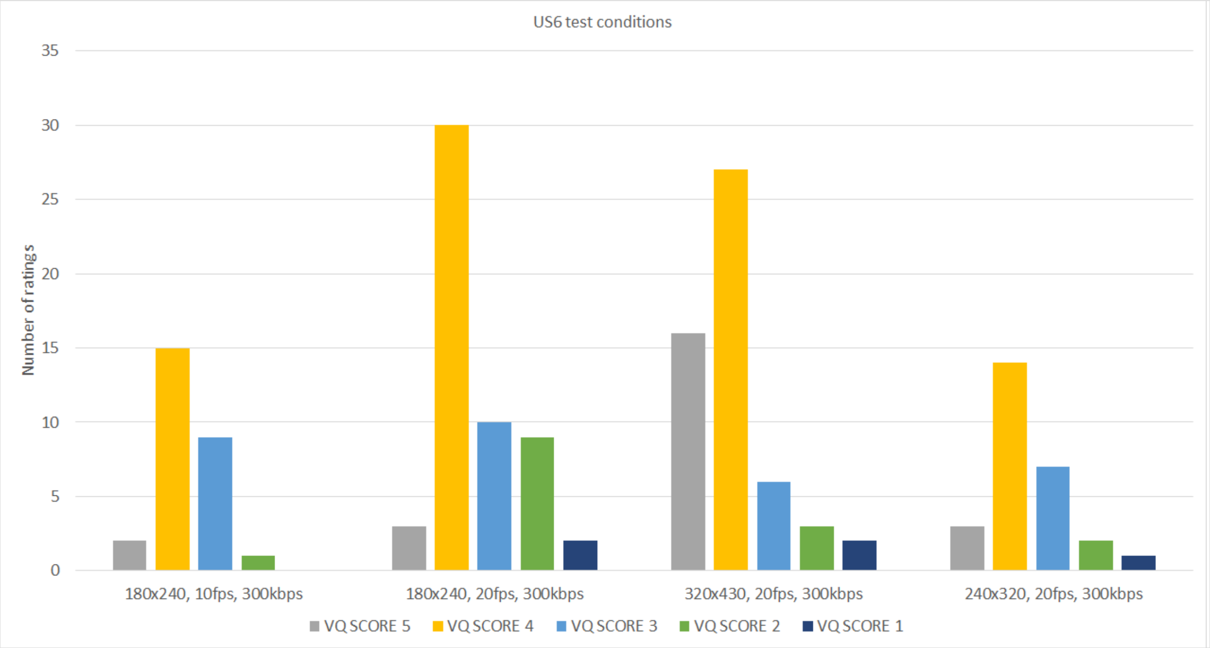


Figure 7.6: Number of reported subjective ratings of video quality for tested conditions in user study US6.

occurred in less than 5% of all sessions per test scenario.

If we take a closer look on the first three VQ MOS results and respective test conditions from Table 7.3 we can calculate Bitrate to Resolution · Frame ratio and show its dependency on the VQ MOS (Figure 7.7).

Factor 0.2 is corresponding to the VQ MOS rating 4.1, which is at the same time average value of the first three VQ MOS results from Table 7.3. Unfortunately there is no one-size-fits-all answer, hence using a factor that is approximately 0.2 (Bitrate = Resolution · Frame rate · 0.2 / 1000 [kbps]) is a good rule of thumb what bitrate per second should the encoder try and target for specific resolution and frame rate to yield good ratings for perceived video quality.

7.1.3 Recommended encoding settings for three-party audiovisual telemeeting in a leisure context

We propose the following general guidelines that can be followed to ensure smooth and well perceived video quality during a three-party video call on smartphone devices:

We recommend streaming at a resolution of 320x480 px with 15 fps and 350 kbps bitrate. Resolutions higher than 320x480 px should only be used when sufficient resources are available (typically high-end smartphones with large displays) to successfully encode and receive video stream without lowering the quality during the video call. Attempting to send higher resolutions via mobile devices with 3GB and 4GB of RAM, can lead to impaired video quality and ultimately poor QoE.

Table 7.3: Acceptability of the video quality per test condition.

Encoding parameters	**Acceptable VQ MOS [%]**	**Unacceptable VQ MOS [%]**	**VQ MOS**
320x480 px, 15 fps, 600 kbps	92.59	7.41	4.18
320x430 px, 20 fps, 300 kbps	77.78	22.22	3.96
320x480 px, 20 fps, 600 kbps	70.37	29.63	4.11
480x640 px, 20 fps, 600 kbps	70.37	29.63	3.92
360x480 px, 15 fps, 300 kbps	66.67	33.33	3.75
240x360 px, 20 fps, 300 kbps	66.67	33.33	3.58
320x480 px, 15 fps, 300 kbps	66.67	33.33	3.78
480x640 px, 15 fps, 600 kbps	62.96	37.04	3.59
180x240 px, 10fps, 300 kbps	62.96	37.04	3.67
240x320 px, 20f ps, 300 kbps	62.96	37.04	3.59
240x360 px, 15 fps, 200 kbps	62.5	37.5	3.67
180x240 px, 20 fps, 300 kbps	59.26	40.74	3.42
180x240 px, 15 fps, 200 kbps	50	50	3.5
240x360 px, 15 fps, 150 kbps	37.5	62.5	3.12
320x480 px, 20 fps, 300 kbps	37.04	62.96	3.41
120x180 px, 20 fps, 200 kbps	33.33	66.67	3.12
480x640 px, 20 fps, 300 kbps	22.22	77.78	2.63
480x640 px, 15 fps, 300 kbps	14.81	85.19	2.71
120x180 px, 15 fps, 100 kbps	4.17	95.83	2.33

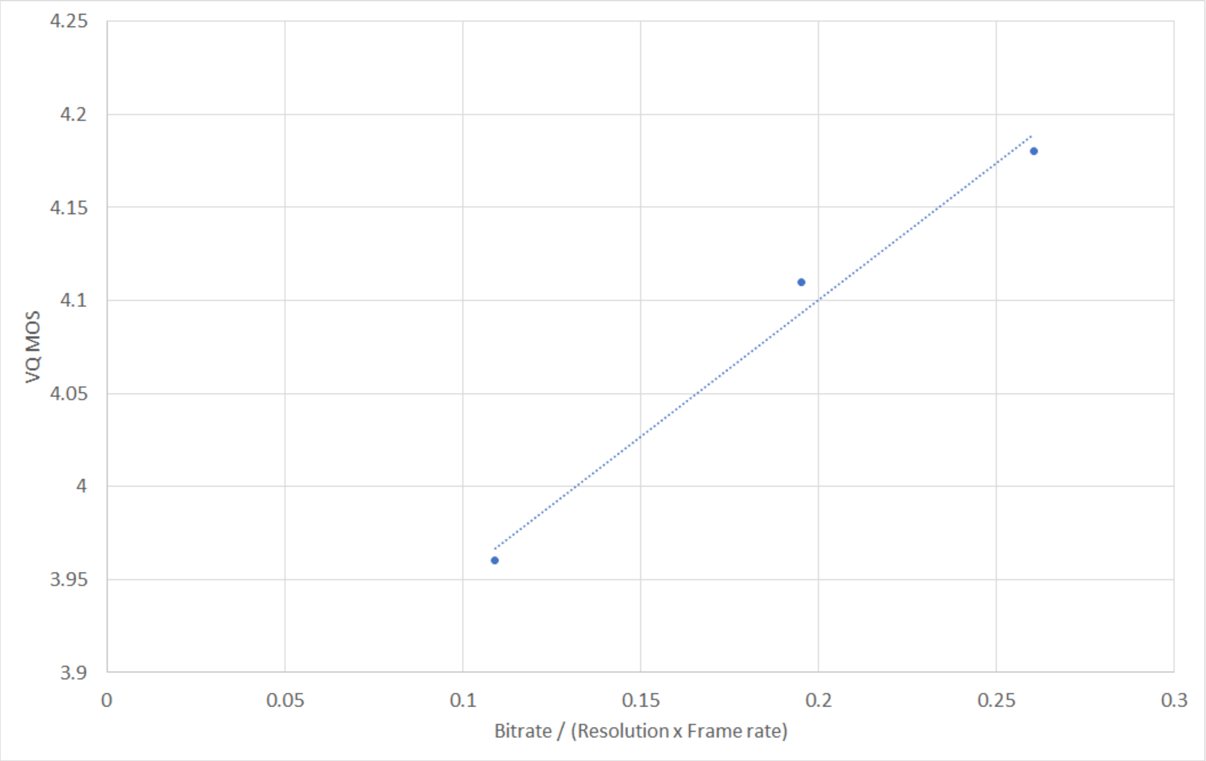


Figure 7.7: Dependency of *Bitrate* to *Resolution* · *Frame rate* ratio on VQ MOS.

Resolution and bitrate should not be increased above 480x640 px and 600 kbps per participant, respectively. More resources than necessary will be used, higher resolutions will require greater processing power to encode the stream, and at the end there will likely be no gain in perceived quality. Many end users viewing on typical smartphone devices will not be able to receive streams at the preset quality level.

7.2 Video encoding adaptation strategies

The ever-increasing demands of users related to being connected anywhere and anytime will force stakeholders to employ new strategies designed to enhance perceived quality. Strategies based on adaptation of video encoding parameters can be proactive or reactive. Proactive adaptation strategies are designed to anticipate possible incoming challenges, while reactive adaptation strategies respond to unanticipated events after their occurrence. Proactive implies environment and context scanning in terms of constant parameters, such as end user device capabilities, while reactive implies in response to dynamic network conditions. A reactive mechanism related to congestion control is already implemented in the WebRTC project in the scope of the Google Congestion Control algorithm [152]. Our focus, on the other hand, is on proposing a complementary, proactive approach aimed to set those codec settings that result in satisfactory QoE, while at the same time attempting to avoid situations that would results in the need to trigger the GCC algorithm. Such an adaptation strategy will be described in the following sections.

Results obtained in conducted user studies served as input for specifying QoE-driven video encoding adaptation strategies, to be triggered in light of system and/or network resource limitations. The goal is to find the preferred resolution to bitrate ratio, which depends on processing capabilities of each display device, movement of the camera or participant, and the available bandwidth. We note that derived QoE models and video adaptation strategies were built solely based on user rating data collected from sessions where video streams were encoded using the VP8 codec. In future studies, the impact of using different codecs should be investigated by repeating user studies so as to investigate the impact of the codec performances on perceived quality. However, established methodologies throughout this thesis, as well as methods for analyzing obtained results and deriving adaptation strategies, may be utilized in future studies involving different codec scenarios.

The following sections present different approaches that could be adopted to formulate the adaptation strategy. Proposed strategies target resolutions up to 480x640 px and 600 kbps, as we found these values during the experiments to be an upper bound for the tested smartphone devices in terms of processing capabilities and perceived differences.

7.2.1 Video adaptation strategy based on derived models

Considering derived QoE and perceived quality models described in Chapter 5, an approach to adaptation strategy formulation has been established. The approach conforms to the concept of QoE maximization, where video bitrate has to be configured in accordance with available uplink video bandwidth. Available bandwidth denotes the available resources for target video bitrates. Parameters should be preset so as to maximize QoE according to the model for perceived video quality (Equation 5.4):

$$PVQ = \frac{-116.723}{VBR} - \frac{15.775}{R} - 0.023 \cdot FR + 4.653$$

Frame rate was the parameter which did not impact greatly perceived video quality, and obtained results showed that 15 fps could provide acceptable QoE. Thus, frame rate could be preset to 15 fps or 20 fps as an upper bound for high movement video calls.

Even though participants were asked to rate features other than video quality, such as perceived audio quality, we did not perform feature in-depth analysis on impairment types. We based our QoE model on perceived audio and video quality (Equation 5.3):

$$QoE = 0.569 \cdot PAQ + 0.43 \cdot PVQ - 0.007$$

The downside of performing adaptation based on the derived QoE model is the resulting frequency of video quality changes invoked in an effort to maintain target QoE, as this may annoy participants. Our studies have shown that it is better to provide lower but constant objective video quality than to switch back and forth between higher and lower qualities. Furthermore, in a mobile context, end user devices often have limited processing capabilities. Hence, adaptation should conform to such limitations as well. Another approach based on the predefined video quality levels emerges from those findings. For each video quality level, the predefined set of video encoding parameters (in terms of bitrate, resolution and frame rate) provides a path to the target QoE level.

7.2.2 Video adaptation strategy based on predefined video quality levels

The idea behind an adaptation strategy based on predefined quality levels is to enable rapid switching between three video quality levels (VQL), depending on available network resources. The VQLs are determined in accordance with tested end user device capabilities, and correspond to high, medium, and low quality level. We propose these VQLs based on the results of our user studies. The high VQL presents video encoding parameters that are able to deliver at least good QoE (referring to parameter settings that resulted with MOS ratings of 4 or higher in our user studies) under typical motion levels during the telemeeting. The medium video quality level should be able to deliver at least fair QoE (referring to parameter settings that resulted with MOS ratings of 3 or higher in our user studies) gravitating toward good QoE, while the lowest VQL present the lower boundary, which should still be able to deliver fair QoE (based on our finding reported in Sections 5.3 and 5.4.2). Video quality under lowest VQL is considered as insufficient and not capable to ensure acceptable perceived quality. We once again note that these quality levels are specified for smartphones with at least 3 GB of RAM and 5.1" display size with three participants previewed simultaneously on the screen, and assuming the video call being held in a leisure context:

- High VQL: 350 kbps, 320x480 px, 15 fps;
- Medium VQL: 250 kbps, 240x360 px, 15 fps;
- Low VQL: 150 kbps, 180x240 px, 15 fps.

This approach is based on available video bandwidth monitoring and estimation at the sender and receiver side, where quality is adapted if estimated bandwidth is insufficient for the currently set quality level (in terms of video coding parameters). The rate controller on the receiver side decides whether to increase, decrease, or hold estimated available bandwidth at the receiver depending on the signal received from the over-use detector (based on the one-way queuing delay variation), while the sender-side has a loss-based controller and calculates (every time an RTCP report or a receiver estimated max bitrate (REMB) message is received from the receiver) the sending bandwidth at the sender depending on the packet loss percentage. The bandwidth value is examined in each cycle (one cycle refers to one second). The session should be started at high VQL, where lower quality levels, if necessary, should be preset so as

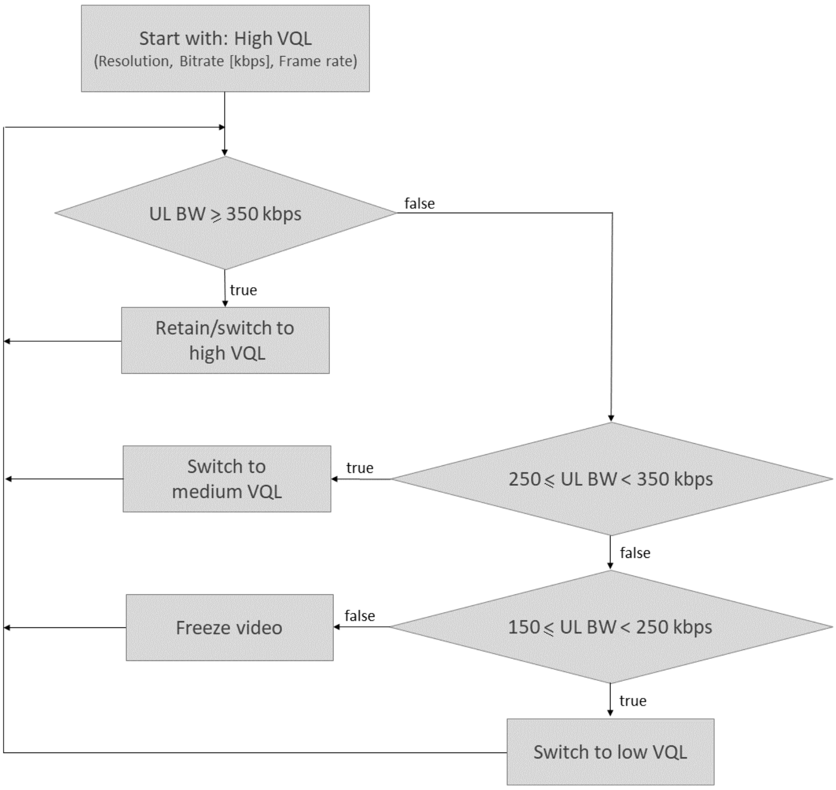


Figure 7.8: Proposed adaptation of video encoding parameters according to three predefined quality levels, namely high, medium and low, in the case of three-party video calls on smartphone devices. The target encoding parameters are set on each sender device according to estimated uplink video bandwidth (UL BW).

to conform to available bandwidth (Figure 7.8).

If the available uplink video bandwidth value falls within the range [150,250>, then we propose for the sent video stream to be encoded with low video quality level, namely bitrate, resolution, and frame rate of 150 kbps, 180x240 px, and 15 fps, respectively. In case when available uplink bandwidth is estimated between 250 kbps and 350 kbps, sent video stream encoding parameters should conform to the medium quality level, with bitrate of 250 kbps, 240x360 px resolution, and 15 fps frame rate. If the available uplink video bandwidth value is equal or greater than 350 kbps, regardless the amount, we propose for the video quality level to be preset at the high level, with encoding parameters configured at 350 kbps, 320x480 px, 15 fps. Results obtained from conducted studies showed that increased bitrate and resolution did not yield higher results in terms of the perceived quality, and as such we determined them as an upper bound in terms of three-party video calls established via smartphones. When the available uplink video bandwidth drops under 150 kbps, the application should freeze the video and stop with the streaming. When estimated uplink bandwidth once again increase beyond 150 kbps, quality can be increased according to the corresponding bitrate. Our proposed algorithm is given in 1.

Algorithm 1 Video encoding adaptation strategy based on three predefined quality levels: high, medium and low

```
Start audiovisual telemeeting on smartphone devices at high video quality level:
Frame width = 320 px
Frame height = 480 px
Bitrate = 350 kbps
Frame rate = 15 fps
while in video call session do
    if estimated uplink video bandwidth $\geqslant$ 350 kbps then
        Retain / switch to high video quality level
    else if 250 kbps $\leqslant$ estimated uplink video bandwidth < 350 kbps then
        Switch to medium video quality level
    else if 150 kbps $\leqslant$ estimated uplink video bandwidth < 250 kbps then
        Switch to low video quality level
    else
        Freeze video
    end if
end while
```

7.2.3 Video adaptation strategy based on CPU usage

The earlier described adaptation approach based on three quality levels was designed in order to conform to a specific three-way setup, where participants are assumed to be using smartphones with at least 3 GB of RAM and approximately 5" screen size. However, when adaptation has to be applied in cases when end user device capabilities are not known in advance, or quality has to be reduced due to the number of participants in a session, we propose to rely on an adaptation strategy based on CPU usage, as described in this section. This adaptation strategy is not targeting any specific quality level, yet aims to determine codec parameters that can result in the best performance given the setup. Adaptation relies on monitoring the CPU level of usage on the sender side, where resolution is lowered when a specific CPU level (e.g., 85%) is exceeded (CPU level usage shall be examined in each cycle - e.g., one second). Considering the encoding resolution has the biggest impact on CPU usage, it is necessary to determine the resolution (for a specific setup and context), and then accordingly the bitrate. Once the limit of the CPU is exceeded, adaptation should be triggered. Sufficient bitrate to support required resolution has to be establish accordingly, trying to deliver best quality image possible that will enable smooth interactivity.

Throughout the course of our research studies, we noticed that a high video quality level (480x640 px, 600kbps and 15 fps) was often reduced by application due to constraints related to smartphone capabilities. We note that the most powerful smartphone that we used for test purposes had 4GB of RAM and eight cores (4x2.3 GHz Mongoose and 4x1.6 GHz Cortex-A53). However, the tremendous development of mobile devices over the past five years led us to examine recent smartphone capabilities. Thus, we reviewed 292 smartphone capabilities and

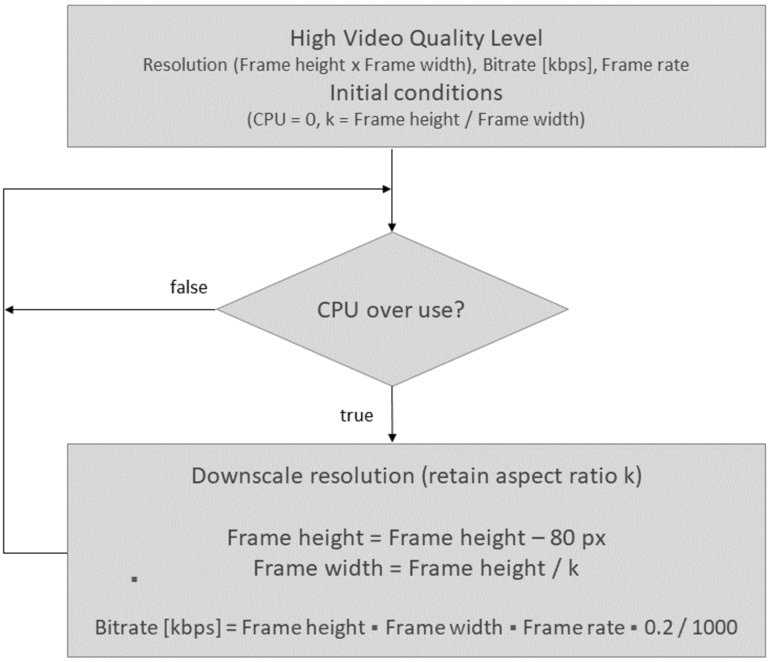

High Video Quality Level
Resolution (Frame height x Frame width), Bitrate [kbps], Frame rate
Initial conditions
(CPU = 0, k = Frame height / Frame width)
CPU over use?
false
true
Downscale resolution (retain aspect ratio k)
Frame height = Frame height – 80 px
Frame width = Frame height / k
Bitrate [kbps] = Frame height ▪ Frame width ▪ Frame rate ▪ 0.2 / 1000

Figure 7.9: Proposed video encoding adaptation strategy based on the CPU usage monitoring.

their corresponding configurations, available on the market between April 2019 and July 2020. Data were taken from websites *GSMArena.com* and *Gadgets.ndtv.com*. The most common value of RAM was 4 GB, and accounted for 26.71% of reviewed smartphones, following by 8 GB smartphones with 23.29%, 6 GB with 21.92%, and 12 GB with 4.45% share. Only 1.71% of reviewed models had 1 GB RAM, while combining 2 and 3 GB of RAM accounted with 21.92% . Taking into consideration availability of more powerful hardware, we based this adaptation approach on the high video quality level of 480x640 px resolution value, and bitrate of 600 kbps. Considering the majority of screen sizes remained under 6 inches, we assume that such high quality level will be adequate for future smartphones as well.

At the beginning of a multiparty video call session established via smartphones should start with highest video quality level (in terms of resolution, bitrate and frame rate: 480x640 px, 600 kbps and 15 fps). If the resolution is too high for the given setup, we propose for it to be downscaled, meaning that frame height should be reduced by 80 px, while preserving the aspect ratio when calculating the frame width (Figure 7.9). This empirically determined value will ensure several video quality levels which are able to support a wide range of smartphone models in terms of hardware capabilities, while the lowest reasonable video resolution is identified as 120x180 px.

When the adequate resolution is determined (e.g., CPU usage measured at the sender device is lower than 85%), target bitrate has to be specified as well, with the goal being to reduce bandwidth consumption with nearly no loss in terms of perceived quality. In a mobile context (referring to the end user mobile device) we prefer CPU over the bandwidth restriction in order to define the feasible resolution first. Algorithm 2 illustrates the process of determining resolution settings considering endpoint processing capabilities in a multiparty setup.

Algorithm 2 Video encoding adaptation strategy based on CPU usage

```
Start audiovisual telemeeting on smartphone device at high video quality level:
Frame width = 480 px
Frame height = 640 px
Bitrate = 600 kbps
Frame rate = 15 fps
k (aspect ratio) = Frame height / Frame width)
while in video call session do
    if CPU usage exceeds certain threshold then
        calculate new resolution and bitrate value
        Frame height = Frame height – 80 px
        Frame width = Frame height / k
        Bitrate [kbps] = Frame height · Frame width · Frame rate · 0.2 / 1000
    else
        retain the video quality level
    end if
end while
```

After the adequate resolution is determined, bitrate should be preset in accordance with the “rule of thumb” (multiplying the resolution and frame rate by 0.2 and dividing by 1000, to get the value in kbps unit). “Rule of thumb” (Section 7.1.2) ensures enough bitrate for smooth interaction with specific motion level of audiovisual telemeeting.

An extended version of the proposed algorithm 2 includes examination of estimated uplink video bandwidth availability and its sufficiency for transmitting a specific video quality level based on the chosen resolution and corresponding bitrate (Figure 7.10). Our proposed algorithm is portrayed in 3.

The two adaptation strategies offer possible ways to address anticipated three-party context with determined lowest smartphone capabilities, while the third adaptation strategy is applicable in a general setup, where it is more difficult for providers to maintain certain QoE. Thus, we proposed an adaptation strategy based on the CPU usage, while the determined video quality level corresponds to the smartphone processing capabilities.

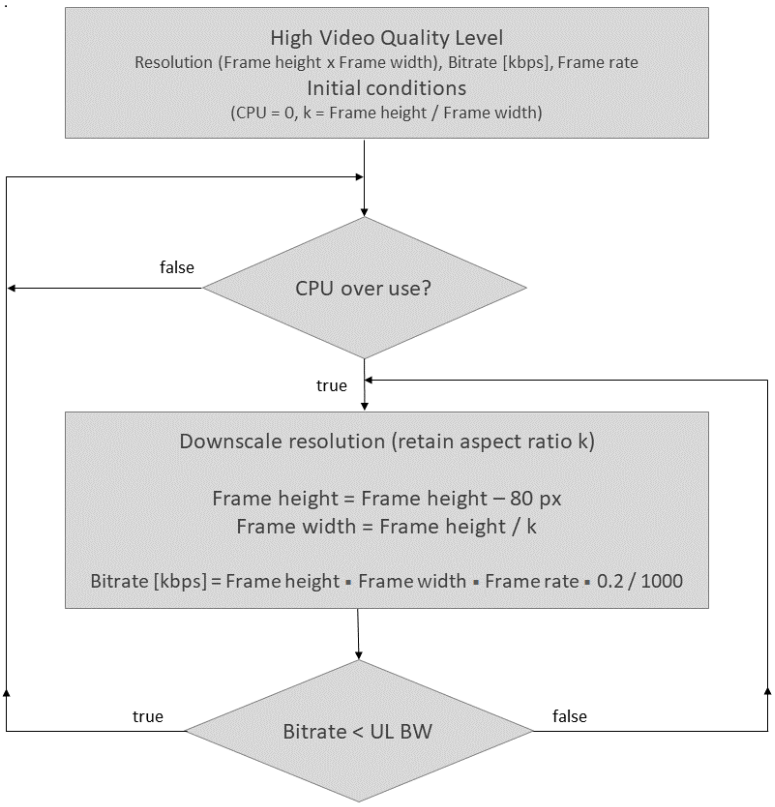

High Video Quality Level
Resolution (Frame height x Frame width), Bitrate [kbps], Frame rate
Initial conditions
(CPU = 0, k = Frame height / Frame width)
false
CPU over use?
true
Downscale resolution (retain aspect ratio k)
Frame height = Frame height – 80 px
Frame width = Frame height / k
Bitrate [kbps] = Frame height ▪ Frame width ▪ Frame rate ▪ 0.2 / 1000
true
Bitrate < UL BW
false

Figure 7.10: Proposed video encoding adaptation strategy based on the CPU usage monitoring and estimated uplink video bandwidth (UL BW).

Algorithm 3 Video encoding adaptation strategy based on CPU usage and estimated uplink bandwidth sufficiency

```
Start audiovisual telemeeting on smartphone device at high video quality level:
Frame width = 480 px
Frame height = 640 px
Bitrate = 600 kbps
Frame rate = 15 fps
k (aspect ratio) = Frame height / Frame width)
while in video call session do
    if CPU usage exceeds certain threshold then
        calculate new resolution and bitrate value
        Frame height = Frame height – 80 px
        Frame width = Frame height / k
        Bitrate [kbps] = Frame height · Frame width · Frame rate · 0.2 / 1000
    else if calculated bitrate > estimated uplink video bandwidth then
        return to the resolution downscale
    else
        retain the video quality level
    end if
end while
```

Summary of key findings

Three different video encoding adaptation strategies are proposed:

- A strategy based on derived QoE and perceived video quality estimation models, where video bitrate is set in accordance with available uplink video bandwidth, and resolution is determined according to the perceived video quality model.
- A strategy based on adaptation between predefined quality levels, where specific video quality level is set in accordance with available uplink bandwidth.
- A strategy based on CPU usage monitoring on the sender device, where resolution is determined in accordance with a predefined CPU usage threshold, and bitrate configured according to the "rule of thumb" which we proposed in Section 7.1.2.

7.3 Chapter summary

In this chapter we proposed three adaptation strategies to respond effectively to limited network and system resources as estimated during a three-party video call established via smartphone devices. Different adaptation strategies present possible adaptation actions that can help to optimize resources while achieving acceptable QoE. These approaches serve as a menu of adaptation actions, where service providers select actions best suited to their specific management objectives. The approaches are derived based on the results of the user studies we reported in sections 5.3 and 5.4.2. Two approaches rely on adaptation of video encoding parameters in response to variable video bandwidth availability, while the third includes CPU usage monitoring (as performed on each sender device). Adaptation strategies based on the bandwidth availability are designed assuming a specific setup (three-way call, via 3 or 4 GB of RAM, approx. 5 inch smartphones) and as such is possibly not suitable for smartphone with lower processing capabilities, while the third adaptation strategy is intended for general setup (in terms of processing capabilities and number of participants). All strategies are based on the configuration of the video encoding bitrate, resolution, and frame rate, with the frame rate being fixed in the second and third approach as we previously established that it has the lowest impact on perceived video quality. In the following chapter, we summarize the contributions of the thesis, discuss the impact and possible future work.

Chapter 8

Conclusion and future work

In this chapter we summarize the overall conclusions reflecting the thesis contributions. In Section 8.1, we summarize the main contributions of the paper and answers to the main thesis research questions (specified in Section 1.4). Finally, research limitations and proposed future work related to assessing and estimating QoE for multiparty audiovisual telemeetings on mobile devices are described in Section 8.2.

8.1 Conclusions and summary of contributions

Based on our initial extensive analysis of state of the art work in Section 3, five research questions were identified and addressed throughout the scope of the thesis. We summarize below our main findings related to answering each of the posed RQ.

RQ1: What are the most influential factors in terms of multiparty audiovisual telemeetings?

To address this research question, a survey was conducted (involving 272 participants), aimed to investigate users' opinions and habits while participating in multiparty video calls via mobile devices. Based on the survey results, the twelve factors that participants considered to be most influential were identified as follows (in descending order): speech intelligibility, audio-video synchronization, longer freezes, perceptible audio delay, low battery consumption, image blurriness, price, security in terms of privacy, ease of use, perceptible video delay, uninterrupted interaction, and installation complexity. Additionally, we identified age and gender as the most influential human factors in terms of expectations, and direct perception, while the most influential context IFs are related to the multiparty setup, such as mobility, number of participants, site distribution, and use case.

RQ2: How can the relationship between QoE and selected video encoding parameters (bitrate, resolution, frame rate) be quantified for multiparty audiovisual telemeetings established via smartphone devices?

Subjective users studies (US5 and US6), conducted in a laboratory environment, aimed to investigate the impact of different video encoding parameters under different system constraints (Chapter 5) on QoE and perceived video quality for multiparty audiovisual telemeetings in a leisure context. With respect to video encoding parameters, obtained results showed that bitrate and resolution have significant impact on the perceived video quality, while frame rate changes did not show significant impact (Note: tested frame rates ranged from 10 to 20 fps). To provide acceptable QoE, the optimal resolution to bitrate ratio, based on typical motion levels that we witnessed during video calls, has to be determined, conforming at the same time to smartphone and network resource availability constraints.

We proposed multidimensional models for QoE and perceived video quality estimation for multiparty audiovisual telemeetings. Models were derived based on the data obtained from user studies US5 and US6 (Chapter 5). Validation of the proposed models and proposed video encoding adaptation strategy was based on the cross-validation measuring the prediction accuracy of the model. The means of the predicted and actual quality values appear to be strongly correlated for both QoE and PVQ models. Both models performed well, with key performance metrics measured as the mean absolute percentage error 3.61% and 2.56% for the QoE model and PVQ model, respectively. Taking into account that QoE is a multidimensional concept with a significant number of impact factors and QoE features, especially in the multiparty mobile context, we can consider both models as relevant, whereas stakeholders of interest may utilize the knowledge of these impacts and relations to enhance their services and to improve users' perceived quality.

RQ3: Can perceived video quality for multiparty audiovisual telemeetings on mobile devices be estimated based on objective video quality metrics?

With respect to objective video quality metrics, we note that our focus was on the no-reference metrics blockiness and blurriness. Objective metrics were calculated based on screen recording of participant smartphones during established video calls. Models designed to estimate objective video quality metrics based on video encoding parameters did not yield satisfactory performance, likely due to the interactivity and movement dynamics specific for video calls established over smartphones, where participants tend to hold smartphones in their hand or move around, causing sporadic video artifacts. Although we obtained the models with accuracy of 75% (eq. 5.6) and 84.5% (eq. 5.7) for blurriness and blockiness respectively, due to the inconsistent results collected in conducted studies, further research is need to gain more confidence in given results. Furthermore, we modeled perceived video quality using blockiness and blurriness as predictors. Again, we obtained high model accuracy of 95.3%. However, participants had difficulties distinguishing blockiness and blurriness, possibly due to the small preview screen size and interaction, meaning that these objective video quality metrics in terms of multiparty audiovisual telemeetings on mobile devices are not good metrics to use when evaluating perceived quality.

RQ4: How can video encoding parameters corresponding to multiparty audiovisual telemeeting services established via smartphone devices be configured so as to optimize end user QoE, given limited processing capabilities of end user mobile devices and bandwidth constraints?

Subsequently, based on the gathered information on the cause of the resolution adaptation (lack of smartphone processing capabilities or insufficient bandwidth) we identified the highest resolution which can be encoded within a particular bitrate without being adapted using interpolation with a rational quadratic function. With this curve, we obtained the maximum *Resolution · Frame rate* multiplication that can be transmitted at a certain bitrate without a resolution reduction in the context of three-party audiovisual telemeetings. Based on our studies, we found that a good "rule of thumb" to determine the bitrate needed for a specific resolution and frame rate to yield good ratings for perceived video quality is using a factor that is approximately 0.2 (*Bitrate = Resolution · Frame rate · 0.2 / 1000 [kbps]*).

RQ5: What is the impact of packet loss on QoE for multiparty audiovisual telemeetings established via mobile devices?

To address the final research question, user studies US3 and US4 were conducted in the field and in a laboratory environment aiming to investigate the impact of network impairments, such as packet loss, on perceived video quality (Chapter 6). We note that our studies only involved scenarios where short bursty packet loss (lasting approx. 10 seconds) was inserted once during 3-minute long test sessions. The results have shown that occasional video impairments caused by packet loss did not significantly impact overall perceived quality (however, we note that participants were only engaged in conversation, and were not focused on presenting to each other any particular visual cues).

8.2 Limitations and future work

Implications of different smartphone capabilities and asymmetric setup

At the time when the research corresponding to this thesis was started, smartphones were launched with a single-core and 512 MHz processor. Today, in early 2021, we are witnessing sixteen-core processors with a clock speed higher than 2.8 GHz on the market. Smartphones are definitely closing the gap between desktop computers with each new release and for sure are able to yield excellent QoE for multiparty video calls. Boosting performance is a nice feature, but it comes with more power consumption and higher price. A powerful processor with a poor amount of available RAM will present a bottleneck in overall performance. Multitasking and processes running in the background rely on a sufficient amount of RAM. To avoid the impact of the end user device, we deliberately conducted most of our measurements in a symmetric setup with high end smartphones (3 and 4 GB of RAM). Thus, future QoE studies should consider to include asymmetric setups (representing realistic use cases) and also smartphones ranging in capabilities covering the whole market range (low-medium-high end).

Implications of the number of participants

Another aspect that plays a key role in audiovisual telemeetings is the number of participants and site distribution. More participants imply different group dynamics and interaction, along with the additional burden in terms of media streams that each end user device has to process. In conducted studies, we focused on the three participant setup. While the four-party scenario in mobile context still might be reasonable, for additional participants, due to the display size and the preview window size of each participant, a dynamic layout should be enabled.

Implications of the context

Video calls can be established for conventional business meetings or for the purpose of more flexible private meetings in a leisure context. Each meeting type has different objectives. While business telemeetings typically have specific objectives and a set of tasks that must be completed, telemeetings held in a leisure context typically have the primary objective of experiencing a sense of presence, and nurturing personal relationships. Due to different objectives, the quality level expected by the participants may be different, with participants likely being less critical in the leisure context. As we conducted subjective studies only in the leisure context, future QoE studies should consider to include the business context as well.

Implications of codec performance

All conducted studies were based on the WebRTC technology utilizing the open, royalty-free, video file format WebM. Proposed QoE models and video adaptation strategies were derived based on data collected in user studies where the video streams were encoded using the VP8 codec. Hence, other available video codecs (such as VP9, H.264, and H.265/HEVC) are out of the scope of this thesis. The usage of different codecs may imply different resource utilization and trade-offs between computational complexity and coding efficiency, and ultimately have a different impact on the perceived quality. Therefore, the performance of the newer codecs should be investigated within additional user studies. However, we note that the methodology used in the thesis, with respect to studying the impact of different coding parameters and deriving adaptation strategy approaches, is generic and may be adopted in future studies.

Improvement of derived QoE models

The video encoding parameters bitrate, frame rate, and resolution were considered in the model analysis of perceived video quality for multiparty audiovisual telemeetings. More application and context dependent information could be used to improve the proposed predictive models accuracy. Within the scope of future work, models could be extended with some of the following influence factors and features:

- perceived reduction in ability to interact with other participants during the video call,
- user mobility (e.g., standing still, walking),
- gender and age,
- end user device processing capabilities.

The future of multiparty audiovisual telemeetings

Important findings collected in the conducted studies, and obtained results based on the subjective user feedback (an important driving force for an adaptation strategy) can be utilized by service providers in an effort to optimize resources and provide acceptable QoE to the users.

Technology enhancements are constantly changing consumer trends. Mobile phones, once serving only as a medium for voice communication and texting, have evolved into smartphone - productive tools, impacting our work habits, education, and relationships. Each new generation brought more advanced hardware in terms of memory, processor, camera, and battery cycle. The smartphone display size tended to get larger as well, trying to accommodate higher resolution screens [156]. However, smartphones with high resolution displays impose additional load on the processing unit, particularly on the graphics in order to render high definition images faster. Commercial launches of the fifth generation of mobile networks started to spread in 2020, enabling faster speeds, lower latency, more responsive connections and more reliable connectivity. All of the aforementioned points indicate incredible development in the mobile industry, implying that the majority of recently released smartphones will be able to provide acceptable QoE for multiparty video calls established over mobile networks, needed now in time of physical distancing more than ever before.

Appendix A

End user online survey

Questionnaire on quality aspects of audiovisual calls established over smartphones in a leisure context

Answers and data collected by this questionnaire will be strictly used for scientific research and will not be used for other purposes. Your personal data will remain anonymous and will not be shown or published anywhere. This survey is designed to gather information about your attitudes and expectations about audiovisual calls.

Media quality refers to the quality of the sound (audio) and the image (video) in terms of perceivable impairments (e.g., delay, blurriness). Please answer the following questions regarding how important you consider disturbances experienced during audiovisual calls.

Please note that the following questions apply to calls established via smartphones in a private/leisure context (e.g., calls with friends, relatives, etc.) and not to calls made for business purposes (e.g., for business meetings).

Nowadays, many video conferencing applications offer additional functionalities beyond only audiovisual calls. Please answer the following questions regarding how important you consider additional functionalities to be.

Usability refers to the ease of use of the application, and the extent to which you feel you are able to make audiovisual calls. Please answer the following questions regarding how important you consider usability, reliability and safety features when participating in audiovisual calls.

Table A.1: Questionnaire: General information

Question	**Answer**
1. How old are you?	18-25 26-35 36-45 46-55 more than 55
2. What is your gender?	Female Male
3. What is your education level?	High school degree University degree (bachelor or masters) Higher University degree (PhD)
4. What is your country of origin?	Croatia Bosnia and Herzegovina Other
5. Do you wear glasses/corrective lenses?	Yes No
6. Please indicate which of the following applications you have used? **(Multiple choices are allowed)**	Skype Google Hangouts Viber Whatsapp Appear.in/Whereby Other video communication app
7. How often have you participated in the above listed applications during the last 30 days approximately?	Very Frequently (on a daily basis) Frequently (2-3 times per week) Occasionally (4-7 time per month) Rarely (1-3 time per month) Never

Continued on next page

Table A.1 – continued from previous page

Question	Answer
8. Which of the following devices have you used in the past to make audiovisual calls? (Multiple choices are allowed)	Smartphone Tablet Computer/laptop Other
9. Have you participated in a video call with more than two users?	Yes No

Table A.2: Questionnaire: Media quality

Question	Answer
10. How important do you consider speech intelligibility for overall audiovisual call quality?	Very Important Important Moderately Important Slightly Important Not Important
11. How important do you consider voice naturalness for overall audiovisual call quality?	Very Important Important Moderately Important Slightly Important Not Important
121. How important is uninterrupted interaction for audiovisual call quality?	Very Important Important Moderately Important Slightly Important Not Important

Continued on next page

Table A.2 – continued from previous page

Question	Answer
13. How important do you consider audio-video synchronization for overall audiovisual call quality	Very Important Important Moderately Important Slightly Important Not Important
14. How important is image sharpness for overall audiovisual call quality?	Very Important Important Moderately Important Slightly Important Not Important
15. How important is smooth movement in the video?	Very Important Important Moderately Important Slightly Important Not Important
16. How important is color accuracy?	Very Important Important Moderately Important Slightly Important Not Important
17. To what extent do you consider perceptible audio delay to impact overall audiovisual call quality?	To a great extent To a moderate extent To some extent To a small extent Not at All

Continued on next page

Table A.2 – continued from previous page

Question	Answer
18. To what extent do you consider perceptible video delay to impact overall audiovisual call quality?	To a great extent To a moderate extent To some extent To a small extent Not at All
19. To what extent do you consider image blurriness to impact overall audiovisual call quality?	To a great extent To a moderate extent To some extent To a small extent Not at All
20. To what extent do you consider that short and occasional video freezes (lasting a few seconds) impact video call quality?	To a great extent To a moderate extent To some extent To a small extent Not at All
21. To what extent do you consider that longer video freezes (i.e., longer than 15 seconds) impact overall audiovisual call quality, if the audio quality remains good for the duration of the call?	To a great extent To a moderate extent To some extent To a small extent Not at All

Table A.3: Questionnaire: Functional completeness

Question	Answer
22. How important is file transfer?	Very Important Important Moderately Important Slightly Important Not Important

Continued on next page

Table A.3 – continued from previous page

Question	Answer
23. How important is texting?	Very Important Important Moderately Important Slightly Important Not Important
24. How important is active speaker identification (i.e., the participant who is currently talking is highlighted/marked in some way)?	Very Important Important Moderately Important Slightly Important Not Important
25. How important is applying make-up / filters / overlay items (e.g., hat, mask)?	Very Important Important Moderately Important Slightly Important Not Important
26. How important is adaptive layout (e.g., movable participant's preview window, display zooming)?	Very Important Important Moderately Important Slightly Important Not Important
27. How important is being able to pause the video?	Very Important Important Moderately Important Slightly Important Not Important

Continued on next page

Table A.3 – continued from previous page

Question	Answer
28. How important is audio mute?	Very Important Important Moderately Important Slightly Important Not Important
29. How important is audiovisual call recording functionality	Very Important Important Moderately Important Slightly Important Not Important

Table A.4: Questionnaire: Usability and service quality

Question	Answer
30. How important is browser / device interoperability?	Very Important Important Moderately Important Slightly Important Not Important
31. How important is the duration of connection time when establishing a call?	Very Important Important Moderately Important Slightly Important Not Important
32. How important is ease of use?	Very Important Important Moderately Important Slightly Important Not Important

Continued on next page

Table A.4 – continued from previous page

Question	Answer
33. How important is installation complexity?	Very Important Important Moderately Important Slightly Important Not Important
34. How important is user interface aesthetics?	Very Important Important Moderately Important Slightly Important Not Important
35. How important do you consider reliability of the service (i.e., being able to use the service - audiovisual call - correctly the first time)?	Very Important Important Moderately Important Slightly Important Not Important
36. How important is security in terms of privacy (i.e., information transmitted during the call is encrypted)?	Very Important Important Moderately Important Slightly Important Not Important
37. How important is low battery consumption during the call?	Very Important Important Moderately Important Slightly Important Not Important

Continued on next page

Table A.4 – continued from previous page

Question	Answer
38. How important is low CPU utilization during the call (allowing for the smooth simultaneous use of other applications)?	Very Important Important Moderately Important Slightly Important Not Important
39. How important is noise-free environment?	Very Important Important Moderately Important Slightly Important Not Important
40. How important is service price?	Very Important Important Moderately Important Slightly Important Not Important
41. Would your answers (expectations) be different if used for free or paid services?	Very Important Important Moderately Important Slightly Important Not Important

Appendix B

QoE questionnaire for three-party audiovisual telemeetings

Answers and data collected by this questionnaire will be strictly used for scientific research and will not be used for other purposes. Your personal data will remain anonymous and will not be shown or published anywhere. This survey is designed to gather information about your attitudes and expectations about audiovisual calls.

Table B.1: Questionnaire: General information

Question	Answer
1. Which video conversation applications have you ever used? **Check all that apply.**	Skype Google hangouts Viber WhatsApp Other
2. How often have you participated in above listed applications during last 30 days approximately? **Mark only one box.**	Never Once Two to three times Once a week Daily
3. Your birth year	_

Continued on next page

Table B.1 – continued from previous page

Question	Answer
4. Your gender:	Male Female
5. What is your occupation?	Employed Unemployed
6. Do you wear glasses/corrective lenses?	Yes No
7. Do you have a special knowledge of AV technology or related field?	Yes No
8. Did you participate before in subjective assessment?	Yes No
9. To which extent are you satisfied with used application on following aspects? **(Rated per each test condition)** Mark one box per row.	Rated aspects: Audio quality Video quality AV synchronization Overall quality
10. Did you perceive any visual impairments during the session? **Multiple answers are allowed.**	Blurriness Blockiness Other
11. Was application frozen during the conference?	Not once Once Two times Several times After some time completely
12. IP address and signal strength:	Fill the data

I *(Name Surname)* hereby declare that the details provided above are true and correct to the best of my knowledge and belief.

(In studies where screen recording was included)

Audio will not be recorded during the sessions. Recorded video streams will be stored on the storage of the Faculty of Electrical Engineering and Computing. Video streams will be used only for objective video quality measurement, as a part of statistical data analysis.

I hereby authorize Dunja Vučić for publicly sharing the information provided on this form.

Place and date Signature

Bibliography

[1] “Number of smartphone users worldwide from 2016 to 2021”, available at https://www.statista.com/statistics/330695/number-of-smartphone-users-worldwide.

[2] “The global internet phenomena report covid-19 spotlight”, available at https://www.sandvine.com/phenomena.

[3] “Video conferencing market size, share and industry analysis”, available at https://www.fortunebusinessinsights.com/industry-reports/video-conferencing-market-100293.

[4] “Idc worldwide smartphone shipment forecast by screen size, 2015-2021”, available at www.idc.com.

[5] Vučić, D., Skorin-Kapov, L., “The impact of packet loss and google congestion control on QoE for WebRTC-based mobile multiparty audiovisual telemeetings”, in International Conference on Multimedia Modeling. Springer, 2019, pp. 459–470.

[6] Wac, K., Ickin, S., Hong, J.-H., Janowski, L., Fiedler, M., Dey, A. K., “Studying the experience of mobile applications used in different contexts of daily life”, in Proceedings of the first ACM SIGCOMM workshop on Measurements up the stack, 2011, pp. 7–12.

[7] Hoßfeld, T., Biedermann, S., Schatz, R., Platzer, A., Egger, S., Fiedler, M., “The memory effect and its implications on web QoE modeling”, in 2011 23rd international teletraffic congress (ITC). IEEE, 2011, pp. 103–110.

[8] Hoßfeld, T., Heegaard, P. E., Skorin-Kapov, L., Varela, M., “No silver bullet: QoE metrics, QoE fairness, and user diversity in the context of QoE management”, in 2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX). IEEE, 2017, pp. 1–6.

[9] “ITU-T Rec. P.800 Methods for objective and subjective assessment of quality”, ITU-T, Tech. Rep., 1996.

[10] “ITU-T Rec. P.911 Subjective audiovisual quality assessment methods for multimedia applications”, ITU-T, Tech. Rep., 1998.

[11] “ITU-T Rec. P.920 Interactive test methods for audiovisual communications”, ITU-T, Tech. Rep., 2000.

[12] “ITU-T Rec. P.805 Subjective evaluation of conversational quality”, ITU-T, Tech. Rep., 2007.

[13] “IT-T Rec. P.880 Continuous evaluation of time-varying speech quality”, ITU-T, Tech. Rep., 2004.

[14] “itu-t rec. p.910 subjective video quality assessment methods for multimedia applications”, Tech. Rep.

[15] “ITU-R Rec. BS.1116 Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems”, ITU-R, Tech. Rep., 1997.

[16] “ITU-R Rec. BS.1285 Pre-selection methods for the subjective assessment of small impairments in audio systems”, ITU-R, Tech. Rep., 1997.

[17] “ITU-T Rec. P.1301 Subjective quality evaluation of audio and audiovisual multiparty telemeetings”, ITU-T, Tech. Rep., 2017.

[18] Carlucci, G., De Cicco, L., Holmer, S., Mascolo, S., “Congestion control for web real-time communication”, IEEE/ACM Transactions on Networking, Vol. 25, No. 5, 2017, pp. 2629–2642.

[19] De Cicco, L., Carlucci, G., Mascolo, S., “Experimental investigation of the google congestion control for real-time flows”, in Proceedings of the 2013 ACM SIGCOMM workshop on Future human-centric multimedia networking, 2013, pp. 21–26.

[20] Skowronek, J., “Quality of experience of multiparty conferencing and telemeeting systems”, Doktorski rad, Ph. D. thesis, Technical University of Berlin, 2017.

[21] Skowronek, J., Schoenenberg, K., Berndtsson, G., “Multimedia conferencing and telemeetings”, in Quality of Experience. Springer, 2014, pp. 213–228.

[22] Skowronek, J., Herlinghaus, J., Raake, A., “Quality assessment of asymmetric multiparty telephone conferences: a systematic method from technical degradations to perceived impairments.”, in INTERSPEECH, 2013, pp. 2604–2608.

[23] Yu, C., Xu, Y., Liu, B., Liu, Y., “Can you see me now? a measurement study of mobile video calls”, in IEEE INFOCOM 2014-IEEE Conference on Computer Communications. IEEE, 2014, pp. 1456–1464.

[24] Jana, S., Chan, A., Pande, A., Mohapatra, P., “QoE prediction model for mobile video telephony”, Multimedia Tools and Applications, Vol. 75, No. 13, 2016, pp. 7957–7980.

[25] Koo, J.-O., Jembre, Y. Z., Choi, Y.-J., Li, Z., Pei, T., Komuro, N., Hiroo, S., “Quality assessment of streaming services in mobile devices”, in 2017 International Conference on Information Networking (ICOIN). IEEE, 2017, pp. 695–699.

[26] Jana, S., Pande, A., Chan, A., Mohapatra, P., “Mobile video chat: issues and challenges”, IEEE Communications Magazine, Vol. 51, No. 6, 2013, pp. 144–151.

[27] Schmitt, M., Gunkel, S., Cesar, P., Bulterman, D., “The influence of interactivity patterns on the quality of experience in multi-party video-mediated conversations under symmetric delay conditions”, in Proceedings of the 3rd International Workshop on Socially-aware Multimedia, 2014, pp. 13–16.

[28] Xu, J., Wah, B. W., “Exploiting just-noticeable difference of delays for improving quality of experience in video conferencing”, in Proceedings of the 4th ACM Multimedia Systems Conference, 2013, pp. 238–248.

[29] Ammar, D., De Moor, K., Xie, M., Fiedler, M., Heegaard, P., “Video QoE killer and performance statistics in WebRTC-based video communication”, in 2016 IEEE Sixth International Conference on Communications and Electronics (ICCE). IEEE, 2016, pp. 429–436.

[30] De Moor, K., Arndt, S., Ammar, D., Voigt-Antons, J.-N., Perkis, A., Heegaard, P. E., “Exploring diverse measures for evaluating QoE in the context of WebRTC”, in 2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX). IEEE, 2017, pp. 1–3.

[31] Zinner, T., Hohlfeld, O., Abboud, O., Hoßfeld, T., “Impact of frame rate and resolution on objective QoE metrics”, in 2010 second international workshop on quality of multimedia experience (QoMEX). IEEE, 2010, pp. 29–34.

[32] Vakili, A., Grégoire, J.-C., “QoE management for video conferencing applications”, Computer Networks, Vol. 57, No. 7, 2013, pp. 1726–1738.

[33] Seeling, P., Fitzek, F. H., Ertli, G., Pulipaka, A., Reisslein, M., “Video network traffic and quality comparison of vp8 and h. 264 svc”, in Proceedings of the 3rd workshop on Mobile video delivery, 2010, pp. 33–38.

[34] Nawaz, O., Minhas, T. N., Fiedler, M., “QoE based comparison of h. 264/avc and webm/vp8 in an error-prone wireless network”, in 2017 IFIP/IEEE Symposium on Integrated Network and Service Management (IM). IEEE, 2017, pp. 1005–1010.

[35] Gunkel, S. N., Schmitt, M., Cesar, P., "A QoE study of different stream and layout configurations in video conferencing under limited network conditions", in 2015 Seventh International Workshop on Quality of Multimedia Experience (QoMEX). IEEE, 2015, pp. 1–6.

[36] Junuzovic, S., Inkpen, K., Hegde, R., Zhang, Z., "Towards ideal window layouts for multi-party, gaze-aware desktop videoconferencing", in Proceedings of Graphics Interface 2011. Canadian Human-Computer Communications Society, 2011, pp. 119–126.

[37] Schoenenberg, K., Raake, A., Egger, S., Schatz, R., "On interaction behaviour in telephone conversations under transmission delay", Speech Communication, Vol. 63, 2014, pp. 1–14.

[38] Schoenenberg, K., Raake, A., Koeppe, J., "Why are you so slow? misattribution of transmission delay to attributes of the conversation partner at the far-end", International Journal of Human-Computer Studies, Vol. 72, No. 5, 2014, pp. 477–487.

[39] Vučić, D., Skorin-Kapov, L., "The impact of mobile device factors on QoE for multi-party video conferencing via WebRTC", in 2015 13th international conference on telecommunications (Contel). IEEE, 2015, pp. 1–8.

[40] Vučić, D., Skorin-Kapov, L., Sužnjević, M., "The impact of bandwidth limitations and video resolution size on qoe for WebRTC-based mobile multi-party video conferencing", screen, Vol. 18, 2016, pp. 19.

[41] Vučić, D., Skorin-Kapov, L., "QoE evaluation of WebRTC-based mobile multiparty video calls in light of different video codec settings", in 2019 15th International Conference on Telecommunications (ConTEL). IEEE, 2019, pp. 1–8.

[42] Vučić, D., Skorin-Kapov, L., "QoE assessment of mobile multiparty audiovisual telemeetings", IEEE Access, Vol. 8, 2020, pp. 107 669–107 684.

[43] Vučić, D., Skorin-Kapov, L., "Investigation of the relationship between subjective and objective video quality metrics for multiparty video calls on mobile devices", in 2021 16th International Conference on Telecommunications (ConTEL). IEEE, 2021, pp. 54–61.

[44] "Video Quality Experts Group (VQEG)", 2020, available at https://www.its.bldrdoc.gov/vqeg/vqeg-home.aspx.

[45] "Different WebRTC architectures", available at https://www.callstats.io/blog/webrtc-architectures-explained-in-5-minutes-or-less.

[46] Loreto, S., Romano, S. P., Real-time communication with WebRTC: peer-to-peer in the browser. O'Reilly Media, Inc., 2014.

[47] "Real-time communication for the web", available at webrtc.org.

[48] "Interactive connectivity establishment (ice): A protocol for network address translator (nat) traversal", Request for Comments, Vol. 8445, 2018.

[49] "Datagram transport layer security (dtls) as transport for session traversal utilities for nat (stun)", Request for Comments, Vol. 7350, 2014.

[50] "Traversal using relays around nat (turn): Relay extensions to session traversal utilities for nat (stun)", Request for Comments, Vol. 5766, 2010.

[51] Grigorik, I., High Performance Browser Networking: What every web developer should know about networking and web performance. O'Reilly Media, Inc., 2013.

[52] Burnett, C., Bergkvist, C., Jennings, A., and Narayanan, "Media capture and streams", Tech. Rep., 2020, available at http://dev.w3.org/2011/webrtc/editor/getusermedia.html, 3rd April 2020.

[53] Roy, R. R., Handbook of SDP for Multimedia Session Negotiations: SIP and WebRTC IP Telephony. CRC Press, 2018.

[54] Valin, J., Vos, K., Terriberry, T., Moizard, A., "Rfc 6716: Definition of the opus audio codec", Internet engineering task force (IETF) standard, 2012.

[55] Bankoski, J., Wilkins, P., Xu, Y., "Vp8 data format and decoding guide", in RFC 6386, 2011.

[56] Holmer, S., Lundin, H., G, C., De Cicco, L., Mascolo, S., "A google congestion control algorithm for real-time communication draft-ietf-rmcat-gcc-02", IETF Informational Draft, 2016.

[57] Zhu, X., Pan, P., Ramalho, M., Mena, S., Jones, P., Fu, J., D'Aronco, S., Ganzhorn, C., "Nada: A unified congestion control scheme for real-time media, draft-ietf-rmcat-nada-02", Internet Engineering Task Force, IETF, 2016.

[58] Johansson, I., Sarker, Z., "Self-clocked rate adaptation for multimedia", draft-johansson-rmcat-scream-cc-05 (work in progress), 2015.

[59] "Mobile and tablet internet usage exceeds desktop for first time worldwide", Tech. Rep., available at https://gs.statcounter.com/press/mobile-and-tablet-internet-usage-exceeds-desktop-for-first-time-worldwides.

[60] Favale, T., Soro, F., Trevisan, M., Drago, I., Mellia, M., “Campus traffic and e-learning during covid-19 pandemic”, Computer Networks, 2020, pp. 107290.

[61] Neustaedter, C., Procyk, J., Chua, A., Forghani, A., Pang, C., “Mobile video conferencing for sharing outdoor leisure activities over distance”, Human–Computer Interaction, Vol. 35, No. 2, 2020, pp. 103–142.

[62] Liu, L., Thorp, S. R., Moreno, L., Wells, S. Y., Glassman, L. H., Busch, A. C., Zamora, T., Rodgers, C. S., Allard, C. B., Morland, L. A. *et al.*, “Videoconferencing psychotherapy for veterans with ptsd: Results from a randomized controlled non-inferiority trial”, Journal of telemedicine and telecare, Vol. 26, No. 9, 2020, pp. 507–519.

[63] Paul, L. R., Salmon, C., Sinnarajah, A., Spice, R., “Web-based videoconferencing for rural palliative care consultation with elderly patients at home”, Supportive Care in Cancer, Vol. 27, No. 9, 2019, pp. 3321–3330.

[64] Jain, R., “COVID-19’s Long-term Impact on Remote Work and Learning”, 2020, available at https://www.nojitter.com/technology-trends/covid-19s-long-term-impact-remote-work-and-learning.

[65] Available at https://www.nojitter.com/technology-trends/covid-19s-long-term-impact-remote-work-and-learning.

[66] “Adobe connect”, https://www.adobe.com/products/adobeconnect.html.

[67] “Fuze”, https://www.fuze.com/products/meetings.

[68] “Anymeeting”, https://www.anymeeting.com/.

[69] “Google meet”, https://meet.google.com/.

[70] “Jami”, https://jami.net/.

[71] “Jitsi mmet”, https://meet.jit.si/.

[72] “Lifesize”, https://www.lifesize.com/.

[73] “Gotomeeting”, https://www.gotomeeting.com/.

[74] “Microsoft teams”, https://www.microsoft.com/en-gb/microsoft-365/microsoft-teams/group-chat-software.

[75] “Skype”, https://www.skype.com/en/.

[76] "Skype business", https://www.microsoft.com/hr-hr/microsoft-365/skype-for-business/download-app.

[77] "Starleaf", https://starleaf.com/.

[78] "Trueconf", https://trueconf.com/.

[79] "Videomost", https://www.videomost.com/en/.

[80] "Wizig", https://www.wiziq.com/video-conferencing-software/.

[81] "zoom", https://zoom.us/.

[82] "Bluejeans", https://www.bluejeans.com/.

[83] "Whereby", https://whereby.com/.

[84] "Uberconference", https://www.uberconference.com/.

[85] "ITU-T Rec. P.10/G.100 Vocabulary for performance, quality of service and quality of experience", ITU-T, Tech. Rep., 2017.

[86] "ETSI TR 102 643 V1.0.1 (2009-12), Human Factors (HF); Quality of Experience (QoE) requirements for real-time communication services", Tech. Rep., 2009.

[87] Niedenthal, P. M., Kitayama, S., The heart's eye: Emotional influences in perception and attention. Academic Press, 2013.

[88] Le Callet, P., Möller, S., Perkins, A., "Qualinet white paper on definitions of quality of experience (2012) version 1.2", in Proc. Eur. Netw. Qual. Exp. Multimedia Syst. Services (COST Action IC), 2013, pp. 1–23.

[89] Jumisko-Pyykkö, S., Vainio, T., "Framing the context of use for mobile hci", International journal of mobile human computer interaction (IJMHCI), Vol. 2, No. 4, 2010, pp. 1–28.

[90] Jumisko-Pyykkö, S., "User-centered quality of experience and its evaluation methods for mobile television", Tampere University of Technology, 2011, pp. 12.

[91] Reiter, U., Brunnström, K., De Moor, K., Larabi, M.-C., Pereira, M., Pinheiro, A., You, J., Zgank, A., "Factors influencing quality of experience", in Quality of experience. Springer, 2014, pp. 55–72.

[92] "Most used smartphone screen resolutions in 2019", available at https://deviceatlas.com/blog/most-used-smartphone-screen-resolutions?imz_s=e938v2126uc4lfoop9ktqekt24.

[93] "ITU-T Rec. G.1080 Quality of experience requirements for IPTV services", ITU-T, Tech. Rep., 2008.

[94] "ITU-T Rec. G.114 Transmission systems and media digital systems and networks", ITU-T, Tech. Rep., 2003.

[95] Möller, S., Berger, J., Raake, A., Wältermann, M., Weiss, B., "A new dimension-based framework model for the quality of speech communication services", in 2011 Third International Workshop on Quality of Multimedia Experience. IEEE, 2011, pp. 107–112.

[96] Husić, J. B., Baraković, S., Veispahić, A., "What factors influence the quality of experience for WebRTC video calls?", in 2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO). IEEE, 2017, pp. 428–433.

[97] García, B., Gallego, M., Gortázar, F., Bertolino, A., "Understanding and estimating quality of experience in WebRTC applications", Computing, Vol. 101, No. 11, 2019, pp. 1585–1607.

[98] Berndtsson, G., Folkesson, M., Kulyk, V., "Subjective quality assessment of video conferences and telemeetings", in 2012 19th International Packet Video Workshop (PV). IEEE, 2012, pp. 25–30.

[99] Winkler, S., "Issues in vision modeling for perceptual video quality assessment", Signal processing, Vol. 78, No. 2, 1999, pp. 231–252.

[100] Chikkerur, S., Sundaram, V., Reisslein, M., Karam, L. J., "Objective video quality assessment methods: A classification, review, and performance comparison", IEEE transactions on broadcasting, Vol. 57, No. 2, 2011, pp. 165–182.

[101] Zeng, K., Zhao, T., Rehman, A., Wang, Z., "Characterizing perceptual artifacts in compressed video streams", in Human Vision and Electronic Imaging XIX, Vol. 9014. International Society for Optics and Photonics, 2014, pp. 90140Q.

[102] Silva, A. F. d., Mylène, C. *et al.*, "Perceptual strengths of video impairments that combine blockiness, blurriness, and packet-loss artifacts", Electronic Imaging, Vol. 2018, No. 12, 2018, pp. 234–1.

[103] Ammar, D., De Moor, K., Skorin-Kapov, L., Fiedler, M., Heegaard, P. E., "Exploring the usefulness of machine learning in the context of WebRTC performance estimation", in 2019 IEEE 44th Conference on Local Computer Networks (LCN). IEEE, 2019, pp. 406–413.

[104] Schmitt, M., Redi, J., Bulterman, D., Cesar, P. S., “Towards individual QoE for multiparty video conferencing”, IEEE Transactions on Multimedia, Vol. 20, No. 7, 2017, pp. 1781–1795.

[105] Husić, J. B., Alagić, E., Baraković, S., Mrkaja, M., “The influence of task complexity and duration when testing QoE in WebRTC”, in 2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH). IEEE, 2019, pp. 1–6.

[106] De Moor, K., Arndt, S., Ammar, D., Voigt-Antons, J.-N., Perkis, A., Heegaard, P. E., “Exploring diverse measures for evaluating QoE in the context of WebRTC”, in 2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX). IEEE, 2017, pp. 1–3.

[107] Jekosch, U., Voice and speech quality perception: assessment and evaluation. Springer Science & Business Media, 2006.

[108] Möller, S., Wältermann, M., Garcia, M., “Features of quality of experience”, in : Möller S., Raake A. (eds) Quality of Experience. T-Labs Series in Telecommunication Services. Springer, Cham. https://doi.org/10.1007/978-3-319-02681-7_5, 2014.

[109] “ITU-T Rec. P.806 A subjective quality test methodology using multiple rating scales”, ITU-T, Tech. Rep., 2014.

[110] “ITU-R Rec. BS.1534 Method for the subjective assessment of intermediate quality levels of coding systems”, ITU-R, Tech. Rep., 2003.

[111] “ITU-T Rec. P.1302 Subjective method for simulated conversation tests addressing speech and audiovisual call quality, institution=ITU-T, author=, year=2014, note = ”, Tech. Rep.

[112] “ITU-T Rec. P.1310 Spatial audio meetings quality evaluation”, ITU-T, Tech. Rep., 2017.

[113] “ITU-T Rec. P.1311 Method for determining the intelligibility of multiple concurrent talkers”, ITU-T, Tech. Rep., 2014.

[114] “ITU-R Rec. BT.500 Methodology for the subjective assessment of the quality of television pictures”, ITU-R, Tech. Rep., 2012.

[115] “ITU-R Rec. BT.710 Subjective assessment methods for image quality in high-definition television”, ITU-R, Tech. Rep., 1998.

[116] “ITU-R Rec. BT.1788 Methodology for the subjective assessment of video quality in multimedia applications”, ITU-R, Tech. Rep., 2007.

[117] "ITU-T Rec. P.915 Subjective assessment methods for 3D video quality", ITU-T, Tech. Rep., 2016.

[118] "ITU-T Rec. P.916 MInformation and guidelines for assessing and minimizing visual discomfort and visual fatigue from 3D video", ITU-T, Tech. Rep., 2016.

[119] "ITU-T Rec. P.913 Methods for the subjective assessment of video quality, audio quality and audiovisual quality of Internet video and distribution quality television in any environment", ITU-T, Tech. Rep., 2016.

[120] "ITU-T Rec. P.1305 Effect of delays on telemeeting quality", ITU-T, Tech. Rep., 2016.

[121] "ITU-T Rec. P.1312 Method for the measurement of the communication effectiveness of multiparty telemeetings using task performance", ITU-T, Tech. Rep., 2016.

[122] "ITU-T Rec. G.1011 Multimedia Quality of Service and performance – Generic and user-related aspects", ITU-T, Tech. Rep., 2015.

[123] Union, I., "Itu-t recommendation p. 800.1: Mean opinion score (mos) terminology", International Telecommunication Union, Tech. Rep, 2006.

[124] Hoßfeld, T., Heegaard, P. E., Varela, M., Möller, S., "QoE beyond the mos: an in-depth look at QoE via better metrics and their relation to mos", Quality and User Experience, Vol. 1, No. 1, 2016, pp. 1–23.

[125] Jansen, B., Goodwin, T., Gupta, V., Kuipers, F., Zussman, G., "Performance evaluation of WebRTC-based video conferencing", ACM SIGMETRICS Performance Evaluation Review, Vol. 45, No. 3, 2018, pp. 56–68.

[126] Egger, S., Reichl, P., Schönenberg, K., "Quality of experience and interactivity", in : Möller S., Raake A. (eds) Quality of Experience. T-Labs Series in Telecommunication Services. Springer, Cham. https://doi.org/10.1007/978-3-319-02681-7_11, 2014.

[127] "ITU-T Rec. J.148 Requirements for an objective perceptual multimedia quality model", ITU-T, Tech. Rep., 2003.

[128] Belmudez, B., Möller, S., "Audiovisual quality integration for interactive communications", EURASIP Journal on Audio, Speech, and Music Processing, Vol. 2013, No. 1, 2013, pp. 24.

[129] Reiter, U., You, J., "Estimating perceived audiovisual and multimedia quality—a survey", in IEEE International Symposium on Consumer Electronics (ISCE 2010). IEEE, 2010, pp. 1–6.

[130] Saidi, I., Zhang, L., Barric, V., Déforges, O., “Audiovisual quality study for videotelephony on ip networks”, in MMSP, 2016.

[131] “ITU-T Rec. G.1070 Opinion model for video-telephony applications”, ITU-T, Tech. Rep., 2018.

[132] Rao, N., Maleki, A., Chen, F., Chen, W., Zhang, C., Kaur, N., Haque, A., “Analysis of the effect of QoS on video conferencing QoE”, in 2019 15th International Wireless Communications & Mobile Computing Conference (IWCMC). IEEE, 2019, pp. 1267–1272.

[133] Balihodžić, M., Husić, J. B., Baraković, S., “The influence of system factors on QoE for WebRTC video communication”, in International Symposium on Innovative and Interdisciplinary Applications of Advanced Technologies. Springer, 2020, pp. 255–267.

[134] Dasari, M., Vargas, S., Bhattacharya, A., Balasubramanian, A., Das, S. R., Ferdman, M., “Impact of device performance on mobile internet QoE”, in Proceedings of the Internet Measurement Conference, 2018, pp. 1–7.

[135] Schmitt, M., Bulterman, D. C., Cesar, P. S., “The contrast effect: QoE of mixed video-qualities at the same time”, Quality and User Experience, Vol. 3, No. 1, 2018, pp. 7.

[136] Karadža, A., Husić, J. B., Baraković, S., Nogo, S., “Multidimensional QoE prediction of WebRTC video communication with machine learning”, in International Symposium on Innovative and Interdisciplinary Applications of Advanced Technologies. Springer, 2020, pp. 269–283.

[137] Scott, M. J., Guntuku, S. C., Huan, Y., Lin, W., Ghinea, G., “Modelling human factors in perceptual multimedia quality: On the role of personality and culture”, in Proceedings of the 23rd ACM international conference on Multimedia, 2015, pp. 481–490.

[138] He, J., Sanadhya, S., Vlachou, C., Qiu, L., Kim, K.-H., “An empirical study to improve qoe estimation for skype in enterprise wireless networks”.

[139] Kale, V., Digital Transformation of Enterprise Architecture. CRC Press, 2019.

[140] Coppens-Hofman, M. C., Terband, H., Snik, A. F., Maassen, B. A., “Speech characteristics and intelligibility in adults with mild and moderate intellectual disabilities”, Folia Phoniatrica et Logopaedica, Vol. 68, No. 4, 2016, pp. 175–182.

[141] “ITU-T Rec. P.807 Subjective test methodology for assessing speech intelligibility”, ITU-T, Tech. Rep., 2016.

[142] “ITU-T Rec. P.931 Multimedia communications delay, synchronization and frame rate measurement”, ITU-T, Tech. Rep., 1998.

[143] Okarma, K., “Mobile video quality assessment: A current challenge for combined metrics”, in Modern Trends and Techniques in Computer Science. Springer, 2014, pp. 485–494.

[144] “ISO/IEC 29100:2011 Information technology — Security techniques — Privacy framework”, Tech. Rep., 2011.

[145] Bezerra, C., De Carvalho, A., Borges, D., Barbosa, N., Pontes, J., Tavares, E., “QoE and energy consumption evaluation of adaptive video streaming on mobile device”, in 2017 14th IEEE Annual Consumer Communications & Networking Conference (CCNC). IEEE, 2017, pp. 1–6.

[146] Alvestrand, “Overview: Real time protocols for browser-based applications”, draft-ietf-rtcweb-overview-19, 2017.

[147] Ammar, D., De Moor, K., Heegaard, P. E., Fiedler, M., Xie, M., “Revealing the dark side of WebRTC statistics collected by google chrome”, in Proceedings from Eighth International Conference on Quality of Multimedia Experience, QoMEX 2016. IEEE, 2016.

[148] “IT-T Rec. J.341 Objective perceptual multimedia video quality measurement of HDTV for digital cable television in the presence of a full reference”, ITU-T, Tech. Rep., 2016.

[149] Schmider, E., Ziegler, M., Danay, E., Beyer, L., Bühner, M., “Is it really robust?”, Methodology, 2010.

[150] Snee, R. D., “Validation of regression models: methods and examples”, Technometrics, Vol. 19, No. 4, 1977, pp. 415–428.

[151] Carlucci, G., De Cicco, L., Holmer, S., Mascolo, S., “Analysis and design of the google congestion control for web real-time communication (WebRTC)”, in Proceedings of the 7th International Conference on Multimedia Systems, 2016, pp. 1–12.

[152] Lundin, H., Holmer, S., Alvestrand, H., “A google congestion control algorithm for real-time communication on the world wide web”, IETF Informational Draft, 2012.

[153] Fouladi, S., Emmons, J., Orbay, E., Wu, C., Wahby, R. S., Winstein, K., “Salsify: Low-latency network video through tighter integration between a video codec and a transport protocol”, in 15th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 18), 2018, pp. 267–282.

[154] “Parametrics curves”, available at https://www.cs.helsinki.fi/group/goa/mallinnus/curves/curves.html.

[155] Turk, S., Računarska grafika. Školska knjiga, Zagreb, 1987.

[156] “Smartphone unit shipments worldwide by screen size from 2018 to 2022”, available at https://www.statista.com/statistics/684294/global-smartphone-shipments-by-screen-size/.

List of Figures

List of Tables

Biography

Dunja Vučić is employed at Ericsson Nikola Tesla in Zagreb, Croatia and working as a system developer for 4G/5G telecom networks and products. She received her M.Sc. degree in the field of Information and Communication Technology in April 2004 from Faculty of Electrical Engineering and Computing, University of Zagreb (FER) and finished her postgraduate master programme in the same field in 2011. Two years later, she enrolled in the PhD program at the Faculty, field of Information and Communication Technology, under the supervision of Prof. Lea Skorin-Kapov. In March 2014 she passed her PhD qualifying exam, and in January 2018 defended the doctoral dissertation topic, entitled "Quality of Experience driven optimization for multiparty video calls on mobile devices based on video encoding adaptation strategy". The focus of the research is on investigating the impact of video encoding parameters configuration so as to maximize participant QoE while meeting resource (network and mobile device) availability constraints in terms of multiparty mobile audiovisual telemeeting. Her research is conducted in the scope of activities of The Multimedia Quality of Experience Research Lab (MUEXlab), and projects funded by the Croatian Science Foundation. She has authored six papers published in international journal and conference proceedings.

List of publications

Publications listed represent work incorporated in this thesis.

Journal paper

1. Vučić, D., Skorin-Kapov, L., "QoE assessment of mobile multiparty audiovisual telemeetings", IEEE Access, Vol. 8, 2020, pp. 107669–107684.

Conference papers

1. Vučić, D., Skorin-Kapov, L., "Investigation of the relationship between subjective and objective video quality metrics for multiparty video calls on mobile devices", 2021 16th International Conference on Telecommunications (ConTEL). IEEE, 2021, pp. 54-61
2. Vučić, D., Skorin-Kapov, L., "QoE evaluation of WebRTC-based mobile multiparty video calls in light of different video codec settings", in 2019 15th International Conference on Telecommunications (ConTEL). IEEE, 2019, pp. 1–8.

3. Vučić, D., Skorin-Kapov, L., “The impact of packet loss and google congestion control on QoE for WebRTC-based mobile multiparty audiovisual telemeetings”, in International Conference on Multimedia Modeling. Springer, 2019, pp. 459–470.

4. Vučić, D., Skorin-Kapov, L., Sužnjević, M., “The impact of bandwidth limitations and video resolution size on QoE for WebRTC-based mobile multi-party video conferencing”, screen, Vol. 18,2016, pp. 19.

5. Vučić, D., Skorin-Kapov, L., “The impact of mobile device factors on QoE for multi-party videoconferencing via WebRTC”, in 2015 13th international conference on telecommunications (Contel). IEEE, 2015, pp. 1–8.

Životopis

Dunja Vučić je zaposlena u Ericsson Nikola Tesla in Zagreb, Hrvatska gdje radi na razvoju sustava za 4G/5G telekomunikacijsku mrežu. Završila je diplomski studij Telekomunikacije i informatika u travnju 2004. godine te 2011. magistarski poslijediplomski studij na istom polju na Fakultetu elektrotehnike i računarstva, Sveučilište u Zagrebu. Dvije godine poslije upisala je doktorski studij, polje telekomunikacije i informatika, pod mentorstvom prof. dr. sc. Lee Skorin-Kapov. U ožujku 2014. položila je kvalifikacijski doktorski ispit, a u siječnju 2018. održala javni razgovor o očekivanom izvornom znanstvenom doprinosu disertacije naslova "Prilagodba kodiranja videa vođena poboljšanjem iskustvene kvalitete višekorisničkih audiovizualnih daljinskih sastanaka na pokretnim uređajima". Fokus istraživanja je na razmatranju utjecaja parametara video kodiranja na iskustvenu kvalitetu korisnika u okviru dostupnih mrežnih resursa i kapaciteta mobilnog uređaja s ciljem maksimiziranja iskustvene kvalitete višekorisničkog video poziva. Istraživanje se provodi u sklopu aktivnosti Laboratorija za istraživanje iskustvene kvalitete višemedijskih usluga (MUEXlab) i projekata financiranih od strane HRZZ-a. Autorica je šest radova objavljenih u međunarodnom časopisu i na konferencijama.