



UNICA

UNIVERSITÀ
DEGLI STUDI
DI CAGLIARI

**Ph.D. DEGREE IN
ELECTRONIC AND COMPUTER ENGINEERING**

Cycle XXXVII

TITLE OF THE Ph.D. THESIS

Advancing Video Communication: From WebRTC Quality Prediction to Green
Applications

Scientific Disciplinary Sector(s)

ING-INF/03

Ph.D. Student: Gülnaziye Bingöl

Supervisor: Prof. Luigi Atzori

Final exam. Academic Year 2023/2024
Thesis defence session: March 2025

This thesis is dedicated to the most important people in my life who have contributed to my growth, learning, and achievements in ways that words cannot fully capture.

To my father, for being my first teacher in life and for instilling in me a deep appreciation for engineering, history, and culture. Your passion for knowledge and your commitment to my education have been a guiding light. I am forever grateful for our Sunday routines to buy books since I was a child, which opened up new worlds for me, and for all the time, effort, and patience you have devoted to shaping my perspective on life. Your wisdom and foresight have been invaluable.

To my mother, for teaching me the power of empathy, patience, art and resilience. Your guidance in developing my soft skills has played a crucial role in shaping the person I am today. Your love, strength, and care have been my foundation and source of comfort through every challenge.

To my sister, for being my lifelong best friend, a constant source of love, joy, and unwavering support. Your friendship has been a source of happiness, and your encouragement has lifted me up in my most difficult moments. Thank you for always being by my side, for making me laugh, and for brightening my days with your energy and warmth.

To Luigi, for your unwavering support and belief in me. Your encouragement has been a source of strength, and your presence has made this journey far more meaningful. Thank you for standing by me every step of the way.

To my supervisors, Prof. Luigi Atzori and Alessandro Floris, PhD for your academic guidance, mentorship, and encouragement in my career development. Your leadership, advice, and support have been pivotal to my journey. Thank you for inspiring me to strive for excellence and for helping me navigate the complexities of academia.

To my friends, for being my anchors and my cheerleaders throughout this journey, especially Sena, Mehmet, Paolo and Chiara. Your companionship has brought joy and balance to my life, and your unwavering belief in me has kept me motivated during the toughest times. Thank you for the laughter, the heartfelt conversations, and for always being there when I needed you most. This journey would not have been the same without your support and encouragement.

Contents

List of Figures	6
List of Tables	8
Abstract	9
1 Introduction	11
1.1 Research Motivation and Context	11
1.2 Research Challenges and Overview	13
1.2.1 Research Questions	13
1.2.2 Research Objectives and Contributions	14
1.2.3 List of Publications	17
1.2.4 Other Publications	18
1.2.5 Awards	18
1.2.6 Outline of the Thesis	19
2 Quality Prediction Solutions for WebRTC	20
2.1 Background and Related work	21
2.2 A Network-Centric Approach	30
2.2.1 Methodology	31
2.2.2 Experimental Results	35
2.2.3 Outcomes and Discussion	38
2.3 Based on Application-layer Data Quality Prediction	39
2.3.1 Statistical Analysis	40
2.3.2 Quality of Experience (QoE) estimation models	44
2.3.3 Outcomes and Discussion	48
2.4 Facial Feature-based Quality Prediction	49
2.4.1 QoE prediction models	49
2.4.2 Performance of QoE prediction models	54
2.4.3 Outcomes and Discussion	57
2.5 Multimodal Quality Prediction	57
2.5.1 Methodology	59
2.5.2 QoE prediction based on the speech features	61

2.5.3	Facial & Speech Fusion: Elevating Quality Predictions	65
2.5.4	Comparison with the state-of-the-art	70
2.6	Key Takeaways and Future Directions	71
3	Balancing Quality and Sustainability in Video Streaming	74
3.1	Energy Consumption of Video Streaming	75
3.1.1	Data Centers and Data Transmission Networks	75
3.1.2	End-device	76
3.1.3	Consumer behavior	77
3.2	Trade-off between Sustainability and QoE	78
3.2.1	Video bitrate vs. Electricity consumption vs. QoE	80
3.3	Are users willing to save energy?	82
3.4	Outcomes and Discussion	83
4	Are Quality and Sustainability Reconcilable? A Case Study on Video Quality, Luminance and Resolution	85
4.1	Scenario Settings	86
4.1.1	Devices and Content Resolution	87
4.1.2	Luminance	89
4.1.3	Electricity consumption	89
4.2	Subjective assessment	92
4.3	Results	93
4.3.1	Mean Opinion Score Analysis	93
4.3.2	Analysis of Variance	94
4.3.3	QoE versus Sustainability Trade-offs	95
4.4	Outcomes and Discussion	96
5	Towards Sustainable 5G: QoE-Driven Energy-Aware Resource Allocation	98
5.1	System Overview	100
5.2	Proposed Solution	101
5.2.1	QoE Modelling	102
5.2.2	QoE and Q based Resource Allocation Algorithm	103
5.3	Performance Evaluation	104
5.3.1	Simulation Settings	104
5.3.2	Results and Analysis	106
5.4	Outcomes and Discussion	108
6	Research Insights and Next Steps	111
6.1	Key Research Findings	111
6.2	Research Challenges and Solutions	113
6.3	Scope and Limitations	114

Appendices	127
A Acronyms	128

List of Figures

2.1	A standard Web Real-Time Communication (WebRTC) setup, signaling between browsers is coordinated by two web servers using proprietary protocols over Hypertext Transfer Protocol (HTTP) or WebSockets.	22
2.2	A secure WebRTC setup where two browsers connect to a web server via HTTP signaling, using a Session Traversal Utilities for NAT (STUN) server for network discovery.	23
2.3	Factors Influencing QoE: System, Context, User, and Behavior Interactions	24
2.4	Mean Opinion Score (MOS) with 95% confidence interval (CI).	31
2.5	The system architecture.	33
2.6	The scenario of the experiment with the participants in different rooms.	33
2.7	MOS with 95% CI for emotions (valence, arousal, and dominance).	37
2.8	Overview of the statistical analysis process for WebRTC session parameters.	41
2.9	Coefficient of determination R^2 results achieved by the 3 regression models when trained on the 3 features datasets.	47
2.10	RMSE results achieved by the 3 regression models when trained on the 3 features datasets.	47
2.11	Variance of the original and synthetic facial expression features' statistics for the ACR scores 1, 2, 4, and 5. Note that a logarithmic scale was used for the y -axis.	52
2.12	FCN architecture. On the top of each box is shown the number of filters used in the convolution step.	54
2.13	The proposed methodology. The dashed lines indicate the collection of ACR scores, which are only needed for the training phase.	60
2.14	The three clusters identified by clustering on the 100 psychoacoustic sharpness features extracted from the recorded speech of the 20 participants. The PCA was used to reduce feature dimensionality to the two principal components (PC1 and PC2).	60

2.15	Variance of the original and synthetic speech features' statistics for ACR scores 1, 2, 4, and 5. Note that a logarithmic scale was used for the y -axis.	64
3.1	QoE (SSIM and MOS) vs. bitrate. The respective values of electricity consumption Q (kWh) are also provided on the graph with colored numbers.	79
3.2	Answers for environmental concern. SD: Strongly Disagree, D: Disagree, NAND: Neither Agree Nor Disagree, A: Agree, SA: Strongly Agree.	83
4.1	Devices' screens with $BL = \text{Min}$. TV: 300 lx, LP: 400 lx, SP: 200 lx. Mean Opinion Score (MOS) with 95% confidence interval (CI) for the first eight TCs (from 1 to 8).	90
4.2	Devices' screens with $BL = \text{Max}$. TV: 5500 lx, LP: 4000 lx, SP: 5000. Mean Opinion Score (MOS) with 95% confidence interval (CI) for the second eight TCs (from 9 to 16).	91
5.1	The considered HetNet deployment scenario.	100
5.2	Average MOS_{TV}^p and Q_{TV} for UF_{TV}^{max} and different α values.	106
5.3	Average MOS_{SP}^p and Q_{SP} for UF_{SP}^{max} and different α values.	106
5.4	Average MOS_{LP}^p and Q_{LP} for UF_{LP}^{max} and different α values.	106
5.5	Average Q for different numbers of devices in the network.	108
5.6	Average MOS for different numbers of devices in the network.	109
5.7	Average BR^{sat} for different numbers of devices in the network.	109

List of Tables

1.1	User Experience Quality Prediction in WebRTC: A Hybrid Framework Combining Network Metrics and Human-centric Features	15
1.2	Energy-efficient Video Streaming: An Adaptive Framework for Quality-aware Management	16
2.1	Overview of QoE Assessment Methods in Video Communication Systems	28
2.2	Test conditions (TCs).	31
2.3	Pearson Correlation Coefficient (PCC) computed between the considered evaluation metrics.	37
2.4	Test conditions (TCs) and MOS.	43
2.5	The hyperparameters for the three ML methods; optimal values are highlighted in bold. MLP used the ADAM optimizer.	45
2.6	The 24 <i>pcc_params</i> . Note: Dec. is the abbreviation for deceleration. .	46
2.7	ANOVA results: features' statistics that achieved a p-value < 0.05 for S_1 , S_2 , and S_3	51
2.8	Number of ACR scores (and corresponding facial features) before and after data augmentation.	53
2.9	QoE estimation performance achieved by the ML-based models and the FCN-based model.	56
2.10	Test conditions (TCs) and MOS with 95% CI.	59
2.11	ANOVA results: LLDs that contain functional statistical features with a p-value < 0.01 for the 3 speech files: OS, NRS, and NSS. . . .	62
2.12	Number of ACR scores (and corresponding speech features) before and after data augmentation for the 3 speech files: OS, NRS, and NSS. .	63
2.13	QoE estimation performance achieved by the ML-based models trained with the speech features.	65
2.14	Number of ACR scores (and corresponding facial and speech features) before and after data augmentation.	66
2.15	QoE estimation performance achieved by the SVM model trained with facial and speech features fused with PCA and ICKA techniques. .	70
2.16	Comparison with the state-of-the-art in terms of the mean accuracy. .	71

3.1	Current possible actions to be taken from the provider's and consumer's side for sustainable video streaming.	76
3.2	End-device data.	78
4.1	The 16 Test conditions (TCs). BL and RES values depend on the end device.	88
4.2	Estimated overall electricity consumption per week for TV, LP, and SP.	90
4.3	Results for the multi-way Analysis of Variance (ANOVA). $p < 0.05$ (*), $p < 0.01$ (**), $p < 0.001$ (***).	93
4.4	T-test results between TCs with $BL = Min$ and TCs with $BL = Max$ while other variables are kept equal. $p < 0.05$ (*), $p < 0.01$ (**), $p < 0.001$ (***).	93
4.5	T-test results between TCs with $RES = High$ and TCs with $RES = Low$ while other variables are kept equal. $p < 0.05$ (*), $p < 0.01$ (**), $p < 0.001$ (***).	94
5.1	Simulation parameters.	105

Abstract

The convergence of Artificial Intelligence (AI) technologies and digital transformation has significantly amplified the importance of understanding and measuring user experience in real-time video communication applications. Therefore, this dissertation presents novel approaches for Quality of Experience (QoE) prediction in Web Real-Time Communication (WebRTC)-based applications through multimodal analysis of naturally available user feedback signals. The research develops comprehensive frameworks for both quality estimation and sustainable service delivery, addressing two critical challenges: the limitations of traditional quality prediction methods and the growing environmental impact of video communications. The first research direction establishes a hybrid framework combining network metrics with human-centric features, specifically facial expressions and speech characteristics, achieving QoE estimation accuracy of 93% through advanced data fusion techniques. The second direction investigates QoE-aware energy efficiency, demonstrating that content characteristics, device configurations, and environmental conditions can be optimized to significantly reduce energy consumption while maintaining acceptable quality thresholds which establishes quantitative frameworks for balancing user experience with environmental impact. The methodological rigor of this work is evidenced through extensive statistical validation, comprehensive experimental design, and thorough performance analysis. The research combines theoretical innovation with practical applicability, providing concrete solutions for implementing energy-aware video communication systems. The frameworks developed enable service providers to make informed decisions about resource allocation, quality management, and energy optimization. The research makes contributions to both QoE prediction methodologies and sustainable video delivery systems, presenting practical solutions for implementing energy-efficient video communications without compromising user satisfaction. The thesis advances the theoretical understanding of QoE evaluation while providing actionable insights for developing user-centric, environmentally conscious multimedia communication systems. In addition, the research outcomes have been validated through extensive experimentation and disseminated through multiple peer-reviewed international conferences and journals. The publication of comprehensive datasets supports reproducibility and facilitates further investigation in this emerging field.

Chapter 1

Introduction

1.1 Research Motivation and Context

The Coronavirus Disease 2019 (COVID-19) pandemic has dramatically accelerated the adoption of video communication technologies, fundamentally transforming how we interact, work, learn, and socialize. Among various video communication solutions, Web Real-Time Communication (WebRTC) has emerged as a pivotal technology, enabling real-time audio and video communication directly through web browsers without requiring additional software installation [1]. This widespread adoption has led to WebRTC-based applications (e.g., Cisco WebEx, Microsoft Teams, and Zoom) becoming integral to modern communication infrastructure, accounting for a significant portion of internet traffic, with video communications representing over 65% of all internet traffic in 2023 [2]. The convergence of AI technologies and digital transformation has significantly amplified the importance of understanding and measuring user experience in real-time video communication applications. As organizations increasingly rely on data-driven decision making and AI-powered service optimization, the ability to accurately assess and predict user experience becomes crucial for maintaining service quality and competitive advantage. This evolution has created a pressing need for innovative approaches that can leverage modern AI capabilities to analyze naturally available user feedback signals, such as facial expressions and speech characteristics, without requiring explicit user intervention or disrupting the communication experience. In this context, QoE, defined as the degree of delight or annoyance of users with an application or service [3], has become a fundamental metric in evaluating communication services. In WebRTC applications, QoE is influenced by various factors, including network conditions (such as delay, jitter, and packet loss), device capabilities, and environmental conditions. Traditional approaches to QoE assessment rely on subjective quality measurements where users explicitly rate their experience. However, this methodology presents several significant limitations: the need for direct user feedback, potential rating scale biases affecting perception measurement accuracy, sub-

stantial time and resource requirements, and most critically, incompatibility with real-time management systems [4]. Researchers have explored various approaches to address these QoE assessment limitations. Initial investigations focused on physiological measurements such as electroencephalography and electrodermal activity [5]. While these methods provide valuable insights into users' perceptual and cognitive processes, their invasive nature and complex implementation render them impractical for real-time quality monitoring in real-world applications. A more promising direction has emerged through affective computing, which analyzes naturally occurring human emotional responses during service interactions. This approach leverages the fundamental connection between user emotions and perceived quality, utilizing Machine Learning (ML) algorithms to automatically interpret behavioral cues such as facial expressions, speech patterns, and body gestures [6], [7]. In WebRTC-based video communications, where users naturally display their faces and engage in conversation, these behavioral signals are readily available without requiring additional user actions or invasive monitoring. By analyzing facial expressions through Action Units (AUs) and speech characteristics through acoustic features, we can develop robust methods for quality estimation that overcome the limitations of traditional approaches while providing accurate, real-time feedback for adaptive system management. While addressing QoE assessment challenges, another critical concern has emerged: the environmental impact of increasing video communication traffic. The Information and Communication Technology (ICT) sector's carbon footprint was estimated at 740 MtCO₂e in 2015, approximately 1.3% of global greenhouse gas emissions, with end-user devices accounting for 54% of these emissions [8]. Video streaming and communication services are substantial contributors to this environmental impact, consuming significant energy across the entire delivery chain - from data centers and network infrastructure to end-user devices. Contemporary research has begun to explore the complex relationship between QoE and sustainability in video communications. Studies have demonstrated that higher service performance does not necessarily correlate with proportionally better perceived quality, suggesting opportunities for energy optimization without significantly impacting user experience. This insight has led to the emergence of "green users" - individuals willing to accept modest sacrifices in QoE to reduce their carbon footprint [9]. The intersection of QoE assessment and sustainability in WebRTC applications presents unique challenges. Unlike traditional streaming services, WebRTC applications require real-time, bidirectional communication with strict latency requirements. These applications must adapt dynamically to changing network conditions while managing energy consumption across heterogeneous devices and network environments. Furthermore, WebRTC's implementation of various error resilience mechanisms, while improving quality, can lead to increased energy consumption in certain scenarios. Through this comprehensive examination of WebRTC video communications, we aim to address both QoE estimation and sustainability challenges. This research is particularly timely as organizations and individuals increasingly rely on video communications while simultaneously facing growing pressure to re-

duce their environmental footprint and enhance user experience through automated, non-intrusive quality assessment methods.

1.2 Research Challenges and Overview

The evolution of WebRTC applications and their widespread adoption presents several significant research challenges that need to be addressed to ensure optimal service delivery and user satisfaction. A primary challenge lies in the limitations of traditional QoE assessment methods, which rely heavily on explicit user feedback and subjective evaluations. These methods, while valuable for certain applications, prove inadequate for real-time monitoring and AI-driven service optimization in modern video communication systems. Current objective QoE models predominantly focus on network parameters, overlooking the wealth of naturally available user-related features that could provide deeper insights into the user experience. The integration of facial expressions and speech characteristics for QoE estimation in WebRTC applications represents another complex challenge. While these features offer promising potential for AI-powered analysis, their effective combination and real-time processing present significant technical hurdles. This includes the need for robust statistical validation of the relationship between extracted features and actual QoE ratings, as well as developing reliable methods for feature selection and data fusion that can scale across diverse usage scenarios. Furthermore, the growing environmental impact of video communication services poses a critical challenge at the intersection of user experience and sustainability. The increasing energy consumption of these applications significantly contributes to ICT's carbon footprint, necessitating careful consideration of the complex interactions between ambient conditions, device settings, and their combined effect on both QoE and energy usage. Understanding and optimizing these relationships while maintaining acceptable user experience across different viewing conditions requires novel approaches to resource management and service delivery.

1.2.1 Research Questions

The challenges identified in video communication systems, particularly regarding QoE estimation and sustainable service delivery, give rise to several fundamental research questions. The increasing complexity of WebRTC applications, combined with the need for real-time quality prediction and environmental sustainability, requires a comprehensive investigation of multiple interconnected aspects of video communication services. The relationship between network conditions, user experience, and video communication quality in WebRTC applications forms the foundation of our initial research direction. Understanding how network impairments affect both service quality and user behavior is crucial for developing effective quality prediction methods. Moreover, the potential of naturally occurring behavioral

signals, such as facial expressions and speech characteristics, presents opportunities for enhanced QoE estimation that warrants thorough investigation. In parallel, the growing environmental impact of video communication services necessitates research into the relationship between QoE and energy consumption. This investigation extends to understanding how various environmental and system parameters interact to influence both user experience and sustainability, leading to critical questions about optimizing service delivery while minimizing environmental impact in video streaming. Guided by these considerations, this research addresses the following specific questions in two main categories:

1. QoE Estimation and Feature Analysis:

- How do network impairments affect conversation quality and perceived user-quality responses in WebRTC applications?
- Which media component (audio or video) demonstrates greater sensitivity to network impairments?
- To what extent can the combination of facial expressions and speech characteristics enhance the QoE estimation accuracy?
- How effective are different data fusion techniques in integrating facial and speech features for robust QoE estimation?

2. Sustainable QoE Management: Impact and Adaptation

- What are the relationships between QoE levels, environmental factors (ambient luminance), system parameters (resolution, content luminance, device backlight) and energy consumption in video streaming?
- How can we optimize the interaction between environmental conditions and device settings to balance user experience and energy efficiency?
- What strategies can be employed to adapt video delivery while considering both environmental conditions and user preferences?

1.2.2 Research Objectives and Contributions

This thesis pursues two complementary research directions: comprehensive statistical analysis of network impairments' impact on QoE in WebRTC applications through feature extraction and modeling, and optimization of video delivery considering both quality and energy consumption aspects. Our objectives are structured based on a systematic progression of research publications, each building upon previous findings to create a comprehensive framework for QoE prediction and sustainable service delivery.

In the first research direction, as shown in Table 1.1, it begins with analyzing how network impairments (delay, jitter, and packet loss) affect conversation quality

Table 1.1: User Experience Quality Prediction in WebRTC: A Hybrid Framework Combining Network Metrics and Human-centric Features

Focus	Key Objectives	Contributions
WebRTC Network Analysis, QoE-1 and QoE-2 [10, 11]	- WebRTC synchronization - Network impairments analysis - WebRTC Application-layer Analysis through ML - Emotional dimensions evaluation	- Audio-visual quality prediction - Network-emotion correlation model - Media component sensitivity analysis
WebRTC Facial-based QoE Estimation, QoE-3 [12]	- Facial expression extraction and modeling - ML-based QoE prediction - Feature significance analysis	- Facial-based QoE - Feature selection methodology
WebRTC Facial-Speech QoE Integration, QoE-4 [13]	- Speech feature analysis - Advanced data fusion design - Enhanced prediction modeling	- Multimodal feature integration - Improved prediction accuracy - Fusion optimization framework
WebRTC QoE Dataset Development [14]	- Feature collection validation - Statistical analysis framework	- Comprehensive feature, network and subjective data - Validation benchmarks

QoE-1, QoE-2, QoE-3, and QoE-4 represent the key contributions of this doctoral thesis in WebRTC service and will be elaborated upon in the following Chapter 2.

and emotional dimensions in WebRTC applications through statistical validation and modeling. The study includes measurement of emotional dimensions (valence, arousal, and dominance) to understand users' emotional responses during impaired conversations. This fundamental analysis is extended to explore facial expression features through AUs extraction and analysis, establishing their significance in QoE estimation. Then, the research progresses to integrate speech characteristics with facial features, employing advanced data fusion techniques to enhance QoE prediction accuracy. This comprehensive analysis culminates in the publication of a validated dataset that encapsulates all investigated features and their relationships with perceived quality. The second research direction, as presented in Table 1.2, focuses on the critical relationship between video streaming quality and sustainability.

Through systematic analysis, we first investigate the fundamental trade-offs be-

Table 1.2: Energy-efficient Video Streaming: An Adaptive Framework for Quality-aware Management

Focus	Key Objectives	Main Contributions
Trade-off between QoE and Sustainability Analysis, QoE-5 [15]	- Energy consumption analysis - CO₂ emission assessment - QoE-Energy tradeoff analysis - User behavior evaluation	- Energy-QoE tradeoff model - User behavior impact analysis - Sustainability framework
QoE-Sustainable Video Quality Assessment, QoE-6 [16]	- Impact of luminance and resolution analysis - Energy consumption measurement - QoE-Sustainability correlation	- Subjective assessment methodology - QoE-aware energy efficiency model
QoE-based Energy-aware Resource Allocation, QoE-7 [17]	- Energy-aware optimization - QoE-based resource management	- Resource allocation framework - Energy efficiency optimization - Network sustainability model

QoE-5, QoE-6, and QoE-7 represent the key contributions of the doctoral thesis in Energy-efficient Video Streaming and will be elaborated upon in the following Chapter 3, 4 and 5.

tween quality requirements and energy consumption in video streaming services. This analysis reveals important insights about user preferences and their willingness to accept quality adjustments for improved sustainability. Building on these findings, we conduct extensive experimental studies examining how different luminance configurations and video resolutions affect both perceived quality and energy consumption across various devices. The research culminates in developing an energy-aware resource allocation framework for Fifth Generation Technology (5G) heterogeneous networks, where we leverage our previous subjective assessment findings to create objective QoE models that effectively balance service quality with energy efficiency. This comprehensive approach enables adaptive optimization strategies that consider both environmental impact and user experience in modern video delivery systems. Building upon these two research directions, the thesis has made contributions to both QoE prediction methodologies and sustainable video delivery systems. Our first direction, detailed in Table 1.1, establishes a comprehensive framework for understanding and predicting user experience through multimodal analysis, combining network metrics, facial expressions, and speech characteristics. This is comple-

mented by our second direction, shown in Table 1.2, which addresses the growing need for sustainable video delivery solutions without compromising user satisfaction. The integration of these research streams has resulted in a holistic approach that not only advances the theoretical understanding of QoE estimation but also provides practical solutions for implementing energy-efficient video delivery systems. Through these contributions, the research of doctoral study have demonstrated that it is possible to achieve a balance between service quality and sustainability, paving the way for future research in user-centric, environmentally conscious multimedia communication systems.

1.2.3 List of Publications

This dissertation's research outcomes have been published in several peer-reviewed international conferences and journals. The thesis is organized based on the following publications:

1.2.3.1 International Journals

- J1. G. Bingöl, S. Porcu, A. Floris, and L. Atzori, "QoE Estimation of WebRTC-based Audio-visual Conversations from Facial and Speech Features," in *ACM Transactions on Multimedia Computing, Communications and Applications (TOMM)*, 2024.

1.2.3.2 International Dataset Journals

- D1. Gülnaziye Bingöl, Simone Porcu, Alessandro Floris, Luigi Atzori, "WebRTC-QoE: A dataset of QoE assessment of subjective scores, network impairments, and facial & speech features", *Computer Networks*, Volume 244, 2024, 110356, ISSN 1389-1286.
- D2. Gülnaziye Bingöl, Luigi Serreli, Simone Porcu, Alessandro Floris, Luigi Atzori, May 18, 2023, "WebRTC-QoE: A Dataset of Quality of Experience in Audio-Video Communications", *IEEE Dataport*, Open Source Dataset [18] available on <https://dx.doi.org/10.21227/mb47-hf44> .

1.2.3.3 International Conferences

- C1. G. Bingöl, L. Serreli, S. Porcu, A. Floris, and L. Atzori, "The Impact of Network Impairments on the QoE of WebRTC Applications: A Subjective Study," in *14th International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 1-6, 2022.
- C2. G. Bingöl, S. Porcu, A. Floris, and L. Atzori, "QoE Estimation of WebRTC-based Audio-visual Conversations from Facial Expressions," in *16th Interna-*

tional Conference on Signal-Image Technology & Internet-Based Systems (SITIS), pp. 577-584, 2022.

- C3. G. Bingöl, S. Porcu, A. Floris, and L. Atzori, "An Analysis of the Trade-off Between Sustainability and Quality of Experience for Video Streaming," in *ICC 2023 - IEEE International Conference on Communications Workshops*, pp. 1600-1605, 2023.
- C4. G. Bingöl, A. Floris, S. Porcu, C. Timmerer, and L. Atzori, "Are Quality and Sustainability Reconcilable? A Subjective Study on Video QoE, Luminance and Resolution," in *15th International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 19-24, 2023. (Diversity and Societal Impact Award , in QoMEX Conference, Ghent/Belgium, 2023)
- C5. M. Hamidi, G. Bingöl, A. Floris, S. Porcu and L. Atzori, "Analysis of Application-layer Data to Estimate the QoE of WebRTC-based Audiovisual Conversations," in *IEEE Globecom Workshops (GC Wkshps)*, pp. 365-370, 2023.
- C6. C. C. González, E. F. Pupo, G. Bingöl, A. Floris, S. Porcu, M. Murrone, and L. Atzori, "A QoE-based Energy-aware Resource Allocation Solution for 5G Heterogeneous Networks," in *16th International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 29-35, 2024.

1.2.4 Other Publications

1.2.4.1 International Conferences

- C1. L. Serreli, G. Bingöl, S. Porcu, A. Floris and M. Martalò, "Robust QUIC-Based Signalling for WebRTC in Impaired Networks," *2023 IEEE International Mediterranean Conference on Communications and Networking (MeditCom)*, Dubrovnik, Croatia, 2023, pp. 187-192,
- C2. F. Loh, G. Bingöl, R. Farahani, A. Pimpinella, R. Prodan and L. Atzori "High Complexity and Bad Quality? Efficiency Assessment for Video QoE Prediction Approaches," *2024 20th International Conference on Network and Service Management (CNSM)*, Prague, Czech Republic, 2024, pp. 1-5

1.2.5 Awards

- Diversity and Societal Impact Award Distinction and Best Paper Finalist at The 15th International Conference on Quality of Multimedia Experience (QoMEX 2023), held in Ghent, Belgium, in recognition of the published work [16].

1.2.6 Outline of the Thesis

The thesis is organized into six chapters that progressively develop the investigation of QoE prediction and optimization in real-time communications, particularly focusing on WebRTC-based systems and sustainable video streaming services through seven major contributions (QoE-1 through QoE-7). Each contribution builds upon previous findings while advancing understanding in their respective domains.

Chapter 1 presents the introduction and research context, establishing the fundamental motivation and objectives of the thesis investigation into QoE estimation methodologies.

Chapter 2 focuses on QoE prediction in WebRTC-based communications, encompassing four major contributions. QoE-1 (Section 2.2) investigates the relationships between network impairments and user experience through subjective assessment, analyzing how delay, jitter, and packet loss affect conversation quality. QoE-2 (Section 2.3) extends this analysis to application-layer data, developing machine learning models for QoE estimation using WebRTC session parameters. QoE-3 (Section 2.4) introduces facial expression-based QoE estimation using machine learning and Fully Convolutional Network (FCN) approaches. QoE-4 (Section 2.5) culminates in the integration of facial and speech features through data fusion techniques, achieving enhanced accuracy in QoE prediction.

Chapter 3 transitions to examining the sustainability aspects of video streaming services. In this chapter, QoE-5 investigates the fundamental trade-offs between energy consumption and user experience in video streaming, identifying the significant impact of end-device usage and consumer behavior on energy costs.

Chapter 4 presents QoE-6, which provides a comprehensive quality evaluation of how video resolution, luminance configurations, and device characteristics affect both QoE and energy consumption in video streaming services. This investigation validates theoretical frameworks with empirical evidence.

Chapter 5 introduces QoE-7, presenting a novel solution for dynamic resource allocation in 5G heterogeneous networks that optimizes the balance between network energy consumption and user QoE. This chapter demonstrates how significant energy savings can be achieved while maintaining acceptable quality levels across different device types.

In the Chapter 6, the thesis concludes by synthesizing these contributions and outlining future research directions, particularly emphasizing the need for deeper investigation of QoE-energy consumption relationships to enable more precise network management approaches that balance service sustainability with user satisfaction.

Throughout the thesis, each chapter maintains rigorous academic standards while progressively building upon previous findings to create a comprehensive framework for understanding and optimizing QoE in modern communication systems. The research demonstrates an advancement in combining traditional network metrics with human-centric features for more accurate quality prediction and sustainable service delivery.

Chapter 2

Quality Prediction Solutions for WebRTC

This chapter provides a comprehensive overview of WebRTC and QoE assessment methodologies in real-time communication systems. The evolution of quality assessment has highlighted the limitations of traditional Quality of Service (QoS) metrics in capturing the full spectrum of user satisfaction. This has led to the emergence of QoE as a critical paradigm in multimedia service management.

The literature review reveals distinct research trajectories in QoE assessment for WebRTC applications, progressing from basic network measurements to a sophisticated analysis of application layer data. However, there is a significant gap in the integration of multiple modalities for comprehensive QoE estimation. This chapter establishes the foundation for the one of thesis's research progression through four key proposals (QoE -1 through QoE -4), each advancing the state-of-the-art in WebRTC QoE assessment through increasingly sophisticated data-driven methodologies.

The sequential progression of this doctoral research on QoE assessment methodologies in WebRTC applications represents a systematically structured investigation that spans across sections 2.1, 2.2, 2.3, 2.4, and 2.5. This comprehensive research trajectory encompasses four distinct yet interconnected phases, each building upon the findings of its predecessors while advancing the theoretical and practical understanding of quality estimation in real-time communications.

The foundational phase, QoE-1 [10] (see in the Section 2.2), establishes the theoretical and empirical groundwork through a meticulous examination of network impairments' influence on WebRTC conversations and their corresponding impact on user emotional responses. This initial investigation provides critical insights into the fundamental relationships between technical network parameters and user experience metrics.

Building upon these foundational insights, QoE-2 [11] (see in the Section 2.3) advances the research through sophisticated analysis of application-layer data, cul-

minating in the development of advanced ML-based QoE estimation models. This phase significantly enhances our understanding of quality prediction mechanisms through computational intelligence approaches.

The research framework is further refined in QoE-3 [12] (see in the Section 2.4), which introduces innovative methodological advancements through the incorporation of facial expression analysis for QoE estimation in WebRTC audiovisual conversations. This phase represents a crucial step toward a more comprehensive quality assessment by integrating human-centric features.

The research culminates in QoE-4 [13] (see in the Section 2.5.3), which presents a sophisticated hybrid framework that seamlessly integrates both facial and speech features for holistic QoE estimation in WebRTC applications. This final phase demonstrates remarkable improvements in prediction accuracy through advanced data fusion techniques, substantially outperforming existing state-of-the-art solutions.

The methodological developments and experimental findings from all four research phases have been meticulously documented and consolidated into a comprehensive dataset, published in [14]. This dataset serves as a resource for the research community, enabling further investigation and advancement in the domain of quality assessment for real-time communications.

2.1 Background and Related work

The rapid growth of multimedia services over the Internet has created unprecedented challenges in managing network resources while meeting increasingly sophisticated user expectations. Video conferencing, streaming platforms, and interactive applications now constitute a significant portion of global Internet traffic, which demands sophisticated approaches to service delivery. Within this landscape, WebRTC has emerged as a transformative open source project that enables real-time communication capabilities directly in web browsers without requiring additional plugins or proprietary software in addition to traditional client-server [1]. WebRTC’s architectural framework follows a model derived from Session Initiation Protocol (SIP) Trapezoid (see Fig. 2.1), which enables web applications from different servers to establish browser-to-browser connections [19, 20]. In a standard WebRTC setup, signaling between browsers is coordinated by two web servers using proprietary protocols over HTTP or WebSockets. After signaling, the browsers create a direct peer-to-peer media connection through the WebRTC PeerConnection Application Programming Interface (API). External services like STUN and Traversal Using Relay around NAT (TURN) servers, represented by the cloud, support secure and reliable communication. Although there are alternative technologies for real-time communication—including SIP for Voice over Internet Protocol (VoIP) systems, Real-Time Messaging Protocol (RTMP) for live broadcasting, and HTTP Live Streaming (HLS) for scalable streaming—WebRTC distinguishes itself through

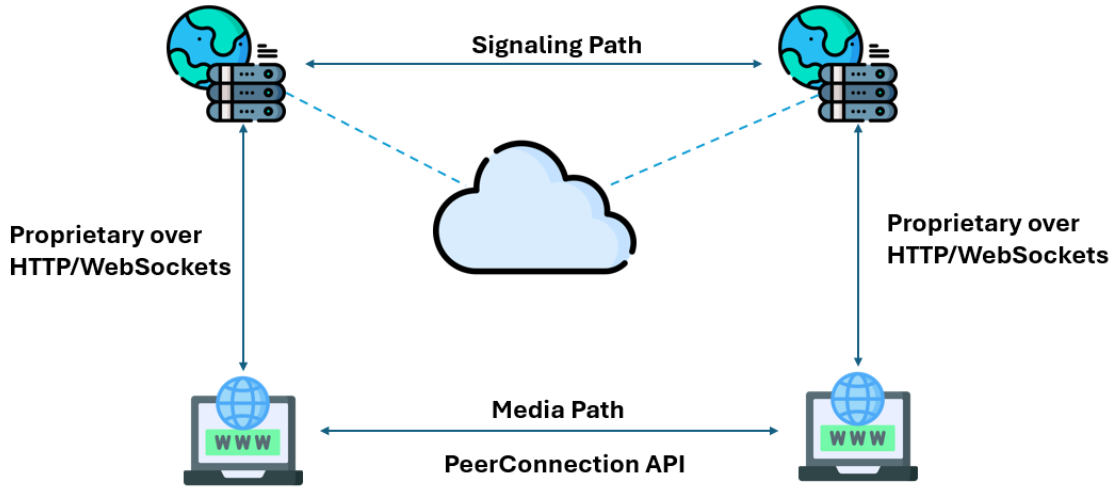


Figure 2.1: A standard WebRTC setup, signaling between browsers is coordinated by two web servers using proprietary protocols over HTTP or WebSockets.

its open source nature and peer-to-peer capabilities, offering advantages over proprietary solutions such as Zoom and Microsoft Teams. The communication architecture employs HTTP or WebSocket protocols for signaling message transmission between browsers and servers, with signaling implementation remaining application-specific rather than standardized [21]. Direct data transfer occurs through PeerConnection, eliminating the need for intermediary servers. In typical deployments, browsers execute identical web applications from a common source, transforming the traditional Trapezoid model into a Triangle configuration, as illustrated in Fig. 2.2. A secure WebRTC setup where two browsers connect to a web server via HTTP signaling, using a STUN server for network discovery. Media communication is relayed through a TURN server to ensure secure and reliable transmission, especially when direct Peer-to-Peer (P2P) connections are blocked by network restrictions. This architectural transition is facilitated by a comprehensive suite of protocols and utilities [22]. Network Address Translation Traversal (NATs) ensures secure message delivery, while Interactive Connectivity Establishment (ICE) enables traversal of firewalls and NATs for peer-to-peer communication. Session Traversal Utilities for NAT (STUN) determines public Internet Protocol (IP) addresses and ports, complemented by Traversal Using Relay around NAT (TURN) servers that provide fallback relay functionality when direct communication paths are unavailable. In practical implementation, this protocol suite operates synergistically. When initiating a video call, NATs facilitates secure message exchange between browsers, followed by ICE negotiation utilizing STUN requests for network address discovery. The system attempts direct peer-to-peer connection establishment while maintaining TURN relay capabilities as a failsafe mechanism for scenarios where network restrictions prevent direct communication [20]. Thus,

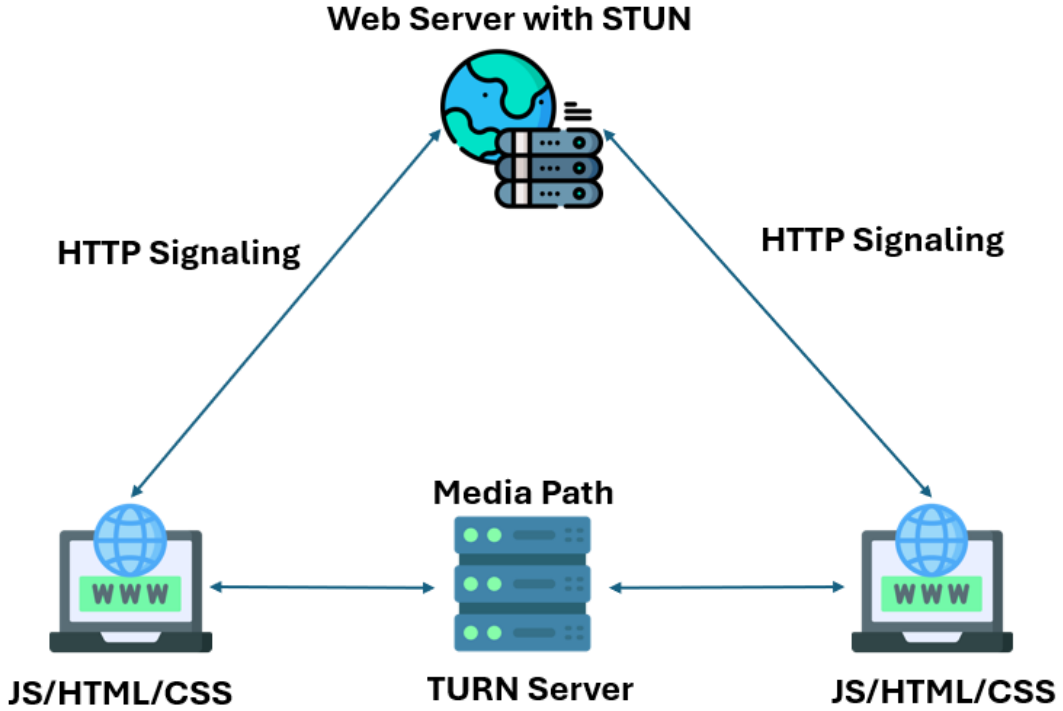


Figure 2.2: A secure WebRTC setup where two browsers connect to a web server via HTTP signaling, using a STUN server for network discovery.

WebRTC has democratized real-time communication development and become the foundation for numerous modern video conferencing and interactive applications.

This evolution has created a pressing need for user-centric management strategies that can effectively balance network resource utilization with user satisfaction. A fundamental challenge in this context is the dynamic allocation of network and server resources to maintain optimal service quality. Traditional QoS metrics alone have proven insufficient to capture the full spectrum of user satisfaction. This has led to the emergence of QoE as a critical paradigm in multimedia service management. QoE assessment provides a comprehensive framework for quantifying user satisfaction, defined as *the degree of delight or annoyance of the user of an application or service* [23]. This user-centric approach has become essential for service providers to optimize resource allocation and maintain competitive advantage in an increasingly demanding market. The evolution of QoE assessment has highlighted the importance of understanding individual user experiences, leading to increased focus on personalized prediction systems [24–26]. This personalization is crucial because QoE perception varies significantly among users, influenced by system, human, and contextual factors [23]. The human factors, particularly objective elements like demographics and cultural background, along with subjective aspects such as enjoy-

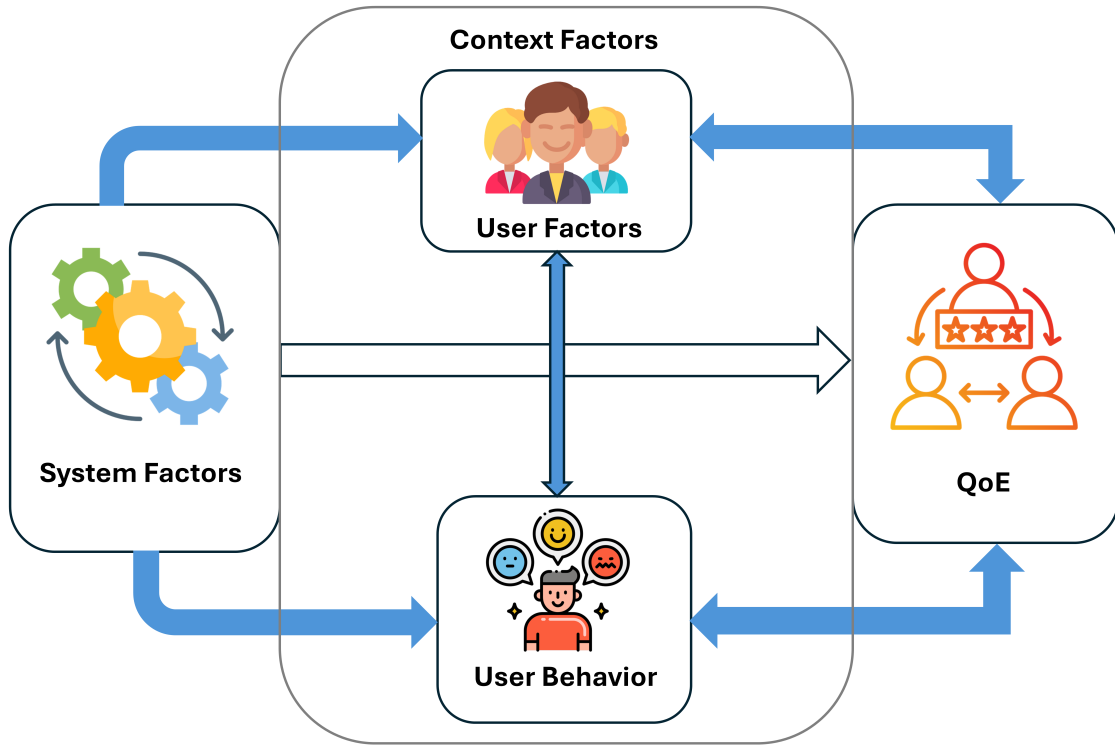


Figure 2.3: Factors Influencing QoE: System, Context, User, and Behavior Interactions

ment and motivation, create distinct user profiles that significantly impact service perception [27], as illustrated in Fig. 2.3.

While personalized QoE models offer improved service management by adapting to individual user profiles [28], traditional subjective assessment methods face several limitations. These include the need for explicit user feedback, potential rating scale biases, resource-intensive data collection, and incompatibility with real-time systems [4].

This has driven research toward unobtrusive assessment methods based on physiological measurements, including electroencephalography [29], heart rate monitoring, and electrodermal activity analysis [30, 31]. Among these approaches, facial expressions and speech characteristics have emerged as particularly promising indicators, as they naturally reflect emotional states and can be captured non-invasively during service usage [32, 33]. These behavioral signals offer a potential solution for real-time QoE assessment while overcoming the limitations of traditional subjective methods. QoE management has evolved through multiple complementary approaches across the service delivery chain [34]. From an application provider perspective, the focus is on end-user level adaptations and cloud service optimizations, particularly evident in adaptive HTTP streaming systems that dynamically respond to network conditions. In contrast, network providers emphasize traffic monitor-

ing and resource optimization within their infrastructure, implementing QoE-driven mechanisms for resource allocation and routing in wireless networks [35]. At the core of these management strategies lies QoE assessment - the ability to accurately predict end-user perceived quality. While substantial research exists for various multimedia services, including video streaming [36–38], VoIP [39], web browsing [40], and virtual reality applications [41], research on WebRTC-based applications remains limited, particularly regarding interactive conversation assessment. The evaluation of quality in audiovisual communication services has traditionally focused on two main approaches: objective quality metrics and passive subjective assessment where participants view and rate pre-recorded content. However, studies involving active participant interaction through real-time audiovisual systems remain limited, particularly for WebRTC-based applications. Early work by Belmudez et al. [42] used the PJSIP VoIP client to study quality during interactive conversations with varying compression and Packet Loss Rate (PLR) (0-5%). Using a building block conversation task taken from the ITU-T Rec. P.920 [43], they found video quality had a stronger influence than audio on overall perceived quality. In [44], the same authors also compared passive versus interactive assessment, finding that users were more sensitive to video impairments during passive viewing compared to interactive conversations where they focused more on the communication exchange. Brauer et al. [45] conducted experiments using Linphone with network delays (100-1800 ms) and packet loss (1-15% audio, 0.1-0.5% video). Interestingly, perceived video quality remained higher than audio quality with delay impairments, with MOS only dropping from 4.50 to 3.46 even at 1800 ms delay. However, packet loss had a more severe impact, particularly on video quality, reducing MOS from 3.68 to 2.32. In examining the packet loss implementations described in [42], [44], and [45], it is important to note that these studies employed a straightforward random packet loss process following the Bernoulli distribution model [46], rather than implementing more complex loss patterns. This methodological choice provides a fundamental approach to simulating network impairments, where each packet has an independent and identical probability of being lost, regardless of the state of previous packet transmissions. The Bernoulli process offers several advantages for experimental design in network impairment studies. First, it provides a controlled and reproducible environment for testing, as the probability of packet loss remains constant throughout the transmission period. Second, this approach allows for clear isolation of the effects of packet loss on QoE, without the additional complexities introduced by burst losses or other pattern-dependent behaviors. However, it should be acknowledged that while this simplified model enables systematic investigation of basic network impairment scenarios, it may not fully capture the complexity of packet loss patterns observed in real-world network conditions, where losses often exhibit temporal correlation and bursty behavior. Nevertheless, the random Bernoulli process remains a valuable baseline methodology for investigating the fundamental relationships between packet loss and perceived quality in WebRTC applications.

More recent work has examined WebRTC applications specifically. De Moor et

al. [47] evaluated four conditions: no impairments, distorted audio (CPU limited), distorted video (20% PLR), and combined audio-video impairments (500 ms delay, 300 ms jitter). Using a name-guessing task, they found combined impairments produced the lowest quality ratings, though emotional valence remained stable. Tsiaras et al. [48] tested delay (150-1600 ms), jitter (0-400 ms) and packet loss (5-40%) to develop their Deterministic QoE Model (DQX) model, which outperformed traditional VoIP quality models like ITU E-model [49] for WebRTC applications. García et al. [50] conducted comprehensive testing across seven network conditions, incorporating packet loss rates up to 45% and jitter values reaching 75ms. Their findings demonstrated that QoE degradation patterns vary across different video content types, with notable quality deterioration occurring at packet loss rates exceeding 20% and jitter beyond 25ms. Building on this work, Serreli et al. [51] demonstrated QUIC's advantages over WebSocket/ Transmission Control Protocol (TCP) for WebRTC signaling under adverse network conditions. Their experiments showed Quick UDP Internet Connections (QUIC) achieves faster connection times and zero error rates even under high packet loss and delay, while WebSocket suffers increasing connection failures. QUIC's architecture, particularly its multi-stream design and improved loss recovery mechanisms, enables more robust performance compared to traditional TCP/WebSocket approaches when network conditions degrade. These findings underscore QUIC's potential for enhancing WebRTC service quality in challenging network environments. However, many earlier studies used dated setups with low resolution video and older codecs, making their findings potentially less applicable to modern WebRTC implementations. Additionally, most work has focused on specific technical parameters rather than comprehensive analysis of network impairments' impact on interactive WebRTC conversation quality. This points to the need for updated research examining quality factors in contemporary WebRTC applications with current video resolutions and codecs.

Building upon the limitations of early VoIP studies, recent research has focused specifically on measuring and modeling WebRTC performance through various approaches. Network-level measurements were pioneered by Yan et al. [52], who explored using QoS parameters (e.g., round-trip time, packet loss, and wireless link signal quality) to predict video freezing and QoE through ML algorithms. While innovative in its ML approach, their work was limited to test video transmission rather than interactive conversations. For multi-party scenarios, researchers in [53–55] conducted extensive studies examining both network-level metrics (bandwidth, packet loss) and application-level parameters (bitrate, frame rate, resolution). Their comprehensive findings demonstrated higher video resolution's positive impact on perceived quality, while highlighting associated resource management challenges. A parallel research direction has leveraged WebRTC's application-level data collected through Google Chrome's *webrtc-internals* tool to develop novel QoE models [56–58]. In [57], researchers demonstrated that correlating performance indicators from both sender and receiver endpoints could enable detection of significant video freezing events. However, their methodology was constrained by limited test scenarios

focusing solely on video transmission without audio components, and the subjective evaluation involved only four participants.

A more comprehensive approach was presented in [56], where researchers utilized the *webrtc-internals* tool to extract features for training multiple ML algorithms. Their study aimed to estimate various quality impairments including audio distortion and video blockiness in multi-party WebRTC conferencing scenarios. While they employed six distinct ML algorithms, their experimental design was limited to four basic test conditions (Good conditions, Video distortion, Audio distortion, and combined Audio-Video distortion). Moreover, their feature selection methodology lacked rigorous statistical validation. In a subsequent study, [58] employed Multiple Linear Regression (MLR) for QoE prediction using *webrtc-internals* statistics. Their experimental framework encompassed 13 distinct scenarios incorporating combinations of three network impairment types: delay, packet loss, and corrupted packets. Through their analysis, they identified four critical *webrtc-internals* metrics as primary QoE predictors: *AudioStreamJitter*, *VideoStreamPacketsSentPerSec*, *VideoStreamPacketsReceivedPerSec*, and *AudioStreamPacketsLost*. The resulting MLR model achieved a coefficient of determination (R^2) of 0.732, indicating reasonable predictive capability.

Beyond traditional network-centric analysis, emerging research has begun exploring human emotional responses as indicators of service quality. Facial expressions, a key component of this approach, can be evaluated through two primary methodologies: the categorical six+one basic emotions framework [59] or the more granular Facial Action Units defined by Facial Action Coding System (FACS) [60]. The latter provides detailed analysis of facial muscle activation patterns and intensities, offering potential insights into emotional states and quality perception. This relationship between emotional expression and quality perception was demonstrated in video conferencing research by Rössler et al. [65]. Their study revealed significant correlations between presenter affect and audience response, with presenter happiness positively influencing audience engagement. Notably, presentations eliciting pronounced emotional responses, particularly alternating between fear and happiness, consistently received higher quality ratings from participants. Recent research has begun exploring QoE estimation models that leverage facial expressions as implicit indicators of user experience. Initial work in this domain focused on HTTP-based video streaming services, where both facial expressions and network QoS parameters were analyzed to estimate QoE. For instance, studies combining facial features with network parameters achieved promising results, demonstrating a Pearson Correlation Coefficient (PCC) of 0.79 between subjective and estimated Mean Opinion Score (MOS) using Random Forest-based approaches [62]. Building on these foundations, more sophisticated approaches emerged incorporating multiple input features. A notable example is found in [7], where facial expressions were combined with gaze direction data and video quality indicators Key Quality Indicator (KQI)s for QoE estimation. This multi-modal approach demonstrated impressive results, with facial and gaze features alone achieving 87.8% accuracy, while

Table 2.1: Overview of QoE Assessment Methods in Video Communication Systems

Work	NB	RT	UB	DT		QoE		WebRTC
				ML	DF	S	O	
[4]	×	×	✓	×	×	×	✓	×
[6]	✓	×	✓	✓	×	✓	✓	×
[7]	✓	×	✓	✓	×	✓	✓	×
[24]	✓	✓	✓	✓	×	✓	✓	×
[26]	✓	×	×	✓	×	×	✓	×
[30]	×	×	✓	×	×	✓	✓	✓
[42]	✓	×	×	×	×	×	×	×
[44]	✓	×	✓	×	×	✓	×	×
[45]	✓	×	×	×	×	✓	×	×
[47]	×	×	✓	×	×	✓	×	✓
[48]	✓	✓	×	×	×	×	✓	×
[50]	✓	✓	×	×	×	✓	✓	✓
[51]	✓	✓	×	×	×	×	✓	✓
[52]	✓	✓	×	✓	✓	×	✓	✓
[53]	✓	✓	✓	×	×	✓	✓	✓
[54]	✓	✓	×	×	×	✓	✓	✓
[55]	✓	×	✓	×	×	✓	✓	✓
[56]	✓	✓	×	×	×	✓	✓	✓
[57]	✓	×	×	×	×	×	✓	✓
[58]	✓	✓	×	×	×	×	✓	✓
[61]	✓	×	×	×	×	✓	✓	×
[62]	✓	✓	✓	✓	×	✓	✓	×
[63]	✓	×	×	×	×	✓	✓	×
[64]	×	×	×	×	×	✓	✓	×
QoE-1 [10]	✓	✓	✓	×	×	✓	×	✓
QoE-2 [11]	✓	✓	×	✓	×	×	✓	✓
QoE-3 [12]	✓	✓	✓	✓	×	✓	✓	✓
QoE-4 [13]	✓	✓	✓	✓	✓	✓	✓	✓

NB: Network Based, RT: Real Time, UB: User Behavior, DT: Data-Driven Techniques, ML: Machine Learning, DF: Data Fusion, S: Subjective O: Objective

the addition of KQI information further improved accuracy to 93.9%. The model showed particularly strong performance when considering simplified quality scales, reaching 97.1% accuracy for 3-level classification and 98.1% for binary classification. Moreover, the strong correlation between subjective and estimated MOS (PCC of 0.989) validated the reliability of facial expression-based approaches.

Speech analysis has also emerged as a powerful tool for understanding user experience and emotional states. Speech features, which encompass both vocal tract features and prosodic features, provide valuable insights into human emotions and quality perception [66, 67]. Vocal tract features, represented by Mel Frequency Cepstral Coefficient (MFCC), characterize different sound units across emotional states. Prosodic features, which include duration, intonation, and intensity, contribute to speech naturalness and are quantified through pitch frequency, duration, and energy-related parameters. These components collectively form Low-Level Descriptor (LLD). Several studies have investigated the relationship between speech features and Quality of Experience (QoE). [63] examined QoE under various network impairments, utilizing acoustic, lexical, and discourse features to train machine learning models. Their Support Vector Machine (SVM) implementation achieved 0.68 accuracy using combined features, with acoustic features demonstrating superior performance compared to lexical and discourse features individually. This superiority stems from acoustic features' capacity to capture speech-related information like pitch and tone, which strongly correlate with emotional states. For instance, fear, anger, and joy typically manifest as loud, rapid speech with elevated pitch ranges, while sadness and fatigue produce slower, lower-pitched speech [26]. Expanding on this foundation, [61] incorporated emotional behaviors alongside vocal and lexical features, achieving 0.828 accuracy with their SVM model—significantly outperforming speech-only features (0.721). The SpeechQoE model [26] further advanced the field by implementing a Convolutional Neural Network (CNN) classifier that processes time-frequency domain spectrograms, achieving 91.4% mean accuracy in QoE estimation across various network conditions. In the context of remote work, [64] demonstrated significant correlations between speech features (MFCCs, Chroma, and Mel features) and perceived quality levels during video calls. ANOVA testing revealed p-values < 0.001 across the Absolute Category Rating (ACR) quality scale. While MFCC and Chroma features effectively distinguished between extreme quality ratings, Mel features demonstrated finer discrimination capability across the quality spectrum, suggesting their potential for training ML-based quality estimation models. However, despite these encouraging results, there remains a significant gap in the literature regarding integrated facial and speech analysis for QoE estimation. Current approaches have typically focused on either facial features for video quality assessment [7, 62] or speech features for audio quality evaluation [26, 63], but rarely combined both. WebRTC-based videoconferencing presents an ideal opportunity to bridge this gap, as it naturally captures both facial expressions and speech during communication. The real-time nature of these applications makes such implicit QoE estimation particularly valuable for dynamic quality optimization

[68] and client-side QoE Management in streaming services [69].

Table 2.1 presents a comprehensive analysis of existing QoE estimation approaches in video/audio communication systems, highlighting the integration of various methodological aspects including network-based measurements, real-time capabilities, user behavior analysis, and data-driven techniques, with particular attention to WebRTC adoption in contemporary solutions.

2.2 A Network-Centric Approach

This section investigates the impact of network impairments on the QoE of WebRTC-based audiovisual conversations with a contribution of the QoE-1 [10]. The current state of research indicates a need for further empirical investigation into how service quality is affected when network conditions are degraded. To address this gap, we first conducted comprehensive subjective tests to collect ground-truth data examining the conversational quality of a two-party WebRTC-based audiovisual telemeeting service. The study aimed to answer three fundamental research questions:

- Q1: What relationships exist between major network impairments and the resulting quality?
- Q2: Which media component demonstrates greater sensitivity to network impairments and to what degree?
- Q3: How do network impairments impact emotional dimensions and communication feasibility?

The experimental assessment involved 20 participants who engaged in conversations through spatially separated rooms using our implemented WebRTC-based application. A dedicated network control system introduced calibrated impairments including delay, jitter, and PLR. Participants evaluated the perceived QoE across three dimensions: audio quality, video quality, and overall service quality. Additionally, they rated three emotional dimensions: valence, arousal, and dominance. Analysis of the results using MOS revealed that audio QoE consistently outperformed video QoE and overall QoE. A notable finding was that the arithmetic mean of audio and video QoE scores served as an effective predictor for overall QoE. The emotional response data indicated that participants maintained relatively pleasant affect throughout the experiment, even during periods of degraded connection quality. The remainder of this section is organized as follows: Section 2.2.1 details our experimental methodology, including system design, test conditions, and assessment procedures. Section 2.2.2 presents a comprehensive analysis of the experimental results. Finally, Section 2.2.3 summarizes key findings and outlines directions for the next step of the research QoE-2 [11].

Table 2.2: Test conditions (TCs).

TC	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Delay (ms)	0	500	1000	500	1000	0	500	1000	500	1000	0	500	1000	500	1000
Jitter (ms)	0	0	0	500	500	0	0	0	500	500	0	0	0	500	500
PLR (%)	0	0	0	0	0	15	15	15	15	15	30	30	30	30	30

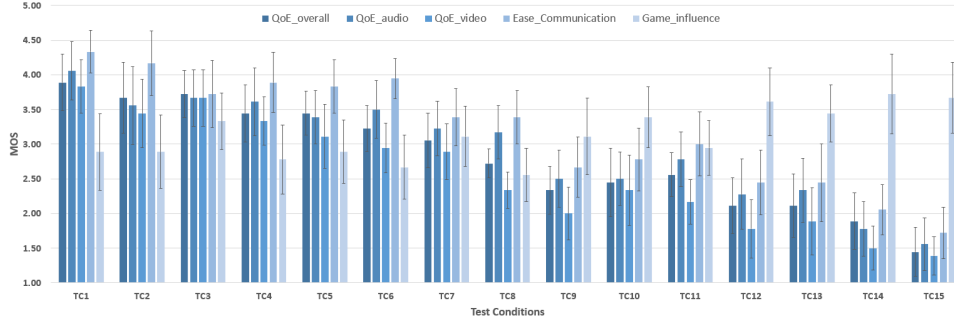


Figure 2.4: Mean Opinion Score (MOS) with 95% confidence interval (CI).

2.2.1 Methodology

2.2.1.1 Experiment Design

The primary objective of this experimental investigation was to assess the conversational quality of a WebRTC-based audiovisual telemeeting service under controlled network impairment conditions. Our methodology focused on three critical network parameters: delay, jitter, and packet loss. These parameters were systematically combined to create 15 distinct test conditions (TCs), as detailed in Table 2.2.

Previous research has identified delay, jitter, and PLR as crucial factors affecting WebRTC video calling applications [57, 70, 71]. Minimal delay is particularly critical for real-time communication scenarios. Jitter, representing variations in packet delay, can induce unintended deviations in both audio and video streams, degrading communication quality. Packet loss, occurring when packets are dropped to mitigate network congestion, significantly impacts audiovisual communication quality [1]. The selected parameter values, while appearing higher than typical VoIP communication thresholds, were deliberately chosen based on WebRTC’s inherent resilience to network impairments. The International Telecommunication Union - Telecommunication Standardization Sector (ITU-T) Rec. G.114 recommends one-way delay remain below 150 ms for VoIP services [72]. Beyond this threshold, perceived QoE degrades exponentially, with delays exceeding 400 ms considered unacceptable [73]. The relationship between QoE and packet loss rate exhibits similar exponential characteristics, influenced by factors such as codec resilience and equipment capabilities. Even modest packet loss rates (e.g., 5%) can render communication quality unacceptable [73]. However, WebRTC applications implement sophisticated error correction mechanisms, including:

- Opus in-band Forward Error Correction (FEC) for audio stream protection
- Redundant audio data (RED), and retransmission mechanisms via Real time Transport Control Protocol (RTCP)
- Google Congestion Control algorithm implementation in Chrome browser for adaptive media rate control [74]

These robust error correction mechanisms render WebRTC applications more resilient to network impairments compared to traditional communication protocols. Consequently, we selected more aggressive impairment values to ensure measurable impact on participant QoE. This approach aligns with previous studies [47, 48] that employed comparable parameter ranges.

2.2.1.2 System Architecture

The experimental system architecture, illustrated in Fig. 2.5, was designed to facilitate controlled network conditions for WebRTC-based audiovisual communications. The fundamental infrastructure comprised two host laptops designated for participant interaction during telemeeting sessions. We implemented the WebRTC-based video call functionality through the Google Chrome browser, with a signaling server deployed on one host laptop to manage WebRTC session synchronization. To establish a controlled experimental environment, we configured a dedicated network topology as depicted in Fig. 2.6. A router facilitated communication between physically separated hosts, creating an isolated network environment that eliminated potential interference from external Internet traffic fluctuations. This configuration was essential for maintaining experimental control over network conditions and ensuring the validity of our impairment measurements. The implementation of precise network impairment controls presented significant technical challenges. While proprietary solutions exist for network traffic manipulation, their substantial cost implications led us to explore open-source alternatives. Our investigation of existing research methodologies revealed successful implementations using various network emulation tools. Notable examples include Dummynet, which has demonstrated efficacy in network protocol testing and simulation [75], and Network Emulator (NetEm) [76], which has been effectively employed to analyze correlations between QoS parameters and QoE in video streaming applications [77]. Our system implementation utilized PyNetem, a sophisticated network emulation framework that establishes a Web server infrastructure for handling dynamic HTTP requests. PyNetem's integration with NetEm rules provided granular control over network conditions while maintaining transparency from the user perspective. However, NetEm's architectural limitation of exclusively affecting outbound traffic [78] necessitated a dual-host deployment strategy, as illustrated in Fig. 2.5. This configuration enabled comprehensive control over both incoming and outgoing traffic streams. To validate the

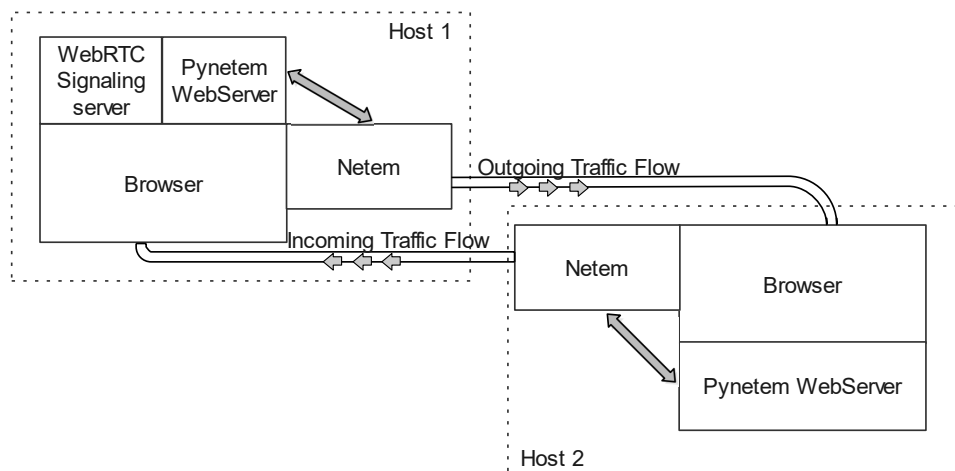


Figure 2.5: The system architecture.

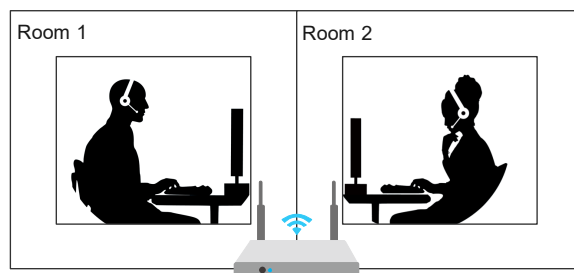


Figure 2.6: The scenario of the experiment with the participants in different rooms.

precision of our network impairment implementation, we employed the iPerf measurement tool for bidirectional verification of network conditions. This rigorous validation process ensured accurate reproduction of our intended test conditions. Following the methodological guidelines established in ITU-T P.800, we implemented randomized assignment of test conditions to participants, effectively minimizing potential parameter correlation effects that could compromise experimental validity. Through this carefully orchestrated system architecture and implementation strategy, we established a robust experimental platform capable of precise network condition manipulation while maintaining ecological validity in the context of real-world WebRTC communication scenarios.

2.2.1.3 Subjective Quality Assessment

The experimental phase involved 20 participants (11 females and 9 males) aged between 23-36 years (mean 28.7, standard deviation 4.24), all of whom reported normal or corrected-to-normal vision and European-Italian origin. For comprehensive demographic characterization, participant skin tones were classified according with five-level [79]. Among male participants, the distribution revealed four in-

dividuals with Type II skin tone classification (light-colored but darker than fair) and five with Type III skin tone classification (golden honey or olive). The female participant distribution showed five individuals with Type II skin tone, five with Type III skin tone, and one with Type IV skin tone (moderate brown). Following the guidelines established in ITU-T Rec. P.920, which emphasizes that vivid audiovisual interactions are best facilitated when participants share prior familiarity, we recruited students from the same academic program who had established interpersonal relationships spanning more than two years [80]. The assessment procedures were conducted at the Department of Electrical and Electronic Engineering (DIEE) of the University of Cagliari, Italy. To maintain experimental control and prevent cross-contamination of audiovisual signals, we implemented a spatially segregated testing environment where participants were placed in separate rooms during their WebRTC sessions. Additionally, participants were required to use headphones throughout the assessment phase to ensure acoustic isolation from external disturbances. For the interactive component of our study, we selected the "Who am I?" celebrity name-guessing paradigm as the conversational task. This choice was informed by ITU-T Rec. P.1305, which suggests that game-based interactions naturally elicit joyful states and spontaneous behavior [81]. The task structure required participants to deduce their assigned celebrity identity through a series of binary (yes/no) questions exchanged with their conversation partner. In accordance with ITU-T Rec. P.920 protocols, participants received comprehensive briefing regarding the system's operational parameters prior to commencing the experimental trials. This included explicit notification that network conditions would vary across telemeeting sessions and that post-session evaluation would require completion of a web-based questionnaire addressing both perceived quality metrics and emotional experience. To establish appropriate calibration benchmarks, participants engaged in three preliminary training sessions demonstrating varying network conditions: optimal performance (no impairments), moderate degradation, and maximum impairment (corresponding to TC15 parameters). The formal experimental phase implemented randomized presentation of all test conditions outlined in Table 2.2. Prior to data collection, participants received thorough instruction in the utilization of multiple rating instruments. The primary quality assessment employed the single discrete ACR scale comprising five categorical levels (Bad, Poor, Fair, Good, and Excellent) for evaluation of overall QoE, video QoE, and audio QoE, as specified by ITU Radiocommunication Sector (ITU-R) Rec. P.800 [82]. Emotional assessment utilized three discrete 5-class graphical rating scales derived from the Self Assessment Manikin (SAM) technique [83], providing non-verbal pictorial evaluation of valence, arousal, and dominance dimensions. Supplementary assessment included two additional metrics: game impact evaluation ("Has the call quality influenced the game?") and communication efficacy assessment ("How easy/difficult was it to communicate with the partner?"), both implemented on 5-point scales ranging from minimal to maximal impact. Each experimental session was temporally constrained to 2 minutes, regardless of task completion status, with participants permitted to

continue their celebrity identification task in subsequent trials if necessary. To mitigate potential fatigue effects, we incorporated a 5-minute rest interval at the experimental midpoint. The complete assessment protocol required approximately 40 minutes per participant pair for full execution of all experimental conditions and associated evaluations.

2.2.2 Experimental Results

This section presents our comprehensive analysis of experimental results derived from 18 test participants. From our initial cohort of 20 participants, two were excluded based on the final rejection criteria for observer discarding as specified in ITU-R BT.500-14 [84]. Despite this reduction, our sample size remains statistically robust, as it exceeds the minimum requirement of 16 participants prescribed by ITU-T Rec. P.920.

2.2.2.1 Analysis of Network Impairments' Impact on Communication Components- Q1 and Q2

Our analysis of the MOS with 95% confidence interval across all 15 test conditions reveals several significant patterns in the relationship between network impairments and perceived quality. As illustrated in Figure 2.4, we examined multiple quality dimensions including overall QoE, audio QoE, video QoE, communication ease, and game task impact. A notable finding emerges regarding audio quality resilience: the audio QoE consistently outperforms both video QoE and overall QoE, with this distinction becoming particularly pronounced under degraded network conditions. Our results demonstrate that participants perceived sufficient to good audio quality (MOS between 3 and 4) in scenarios where PLR was absent (TCs 1-5) or when PLR was limited to 15% either in isolation or combined with delay (TCs 6-8). The introduction of jitter alongside delay and 15% PLR (TCs 9-10) precipitated a decline in audio QoE to poor-sufficient quality range (MOS between 2 and 3). Similar quality degradation was observed with 30% PLR, whether isolated or combined with delay (TCs 11-13). The most severe audio quality deterioration, characterized by poor to bad MOS ratings between 1 and 2, occurred when jitter augmented the combination of delay and 30% PLR (TCs 14-15). Video quality demonstrated greater sensitivity to network impairments, consistently receiving the lowest ratings compared to overall and audio QoE. This disparity became particularly evident under substantial network degradation, especially in the presence of packet loss. Sufficient to good video quality (MOS between 3 and 4) was maintained only in scenarios without packet loss (TCs 1-5). Quality degraded to poor-sufficient levels (MOS between 2 and 3) with 15% PLR regardless of delay and jitter combinations (TCs 6-10), and with isolated 30% PLR (TC 11). The most severe video quality impairment (MOS between 1 and 2) occurred when 30% PLR combined with delay and jitter (TCs 12-15). Our comparison with previous research reveals alignment in single-variable sce-

narios, with delay and PLR measurements corresponding to those reported by [48]. Multi-variable MOS results demonstrate general concordance with [47], though with less pronounced differentiation between test conditions. For instance, while [47] reported significantly lower ratings (MOS approximately 2) for audio-video distortion with 500 ms delay and 300 ms jitter compared to video-only impairment with 20% PLR (MOS approximately 3), our findings show more moderate distinctions. Specifically, TC4 (500 ms delay, 500 ms jitter) and TC6 (15% PLR) achieved MOS values of 3.44 and 3.22 respectively. The aggregate MOS calculations across quality dimensions (overall: 2.80, audio: 2.93, video: 2.57) substantiate the superior resilience of audio quality perception. This enhanced audio robustness can be attributed to two primary factors: the reduced bandwidth requirements of audio transmission compared to video data, and the effectiveness of WebRTC’s implemented concealment techniques in masking network impairments. Furthermore, as noted in [85], the audio stream’s immunity to Google Congestion Control due to its inherently low data rate contributes to this stability. A particularly noteworthy finding emerges in the relationship between overall quality perception and component qualities. The overall MOS demonstrates remarkable proximity to MOS_{AV} , the arithmetic mean of audio and video MOS values. This relationship is quantitatively validated through statistical analysis, yielding a Mean Squared Error of 0.01 and a Pearson Correlation Coefficient of 0.993 between overall MOS and MOS_{AV} , establishing the latter as a reliable predictor of overall quality perception. The ease of communication metric generally exceeded QoE ratings, though following similar degradation patterns with increasing network impairments. This inverse relationship between communication facility and network degradation manifested as expected. However, the game task performance metrics reveal an interesting resilience in basic communication capability. Even under impaired conditions, participants maintained sufficient information exchange for celebrity identification tasks. The impact on game performance only reached significant levels when PLR exceeded 30% in combination with delay and jitter (TCs 12-15). Correlation analysis through PCC calculations (Table 2.3) reveals complex relationships between evaluation metrics. Overall QoE demonstrates strong correlations across all metrics, with the strongest correlation observed with video QoE (0.988) and the weakest (negative) correlation with game task impact (-0.681). While audio QoE shows slightly lower correlation with overall QoE (0.979) compared to video QoE, it exhibits stronger correlations with communication ease (0.986) and game impact (-0.778) than both overall QoE (0.977, -0.681) and video QoE (0.951, -0.617). These correlations reinforce audio quality’s primacy as a critical quality indicator, suggesting that superior audio QoE facilitates both basic communication and task completion.

Table 2.3: Pearson Correlation Coefficient (PCC) computed between the considered evaluation metrics.

PCC	Overall QoE	Audio QoE	Video QoE	Ease of Comm.	Game Influence	Valence	Arousal	Dominance
Overall QoE	1	0.979	0.988	0.977	-0.681	0.980	0.898	0.949
Audio QoE		1	0.962	0.986	-0.778	0.986	0.899	0.957
Video QoE			1	0.951	-0.617	0.961	0.913	0.951
Ease of Comm.				1	-0.801	0.984	0.886	0.949
Game Influence					1	-0.744	-0.610	-0.669
Valence						1	0.914	0.942
Arousal							1	0.930
Dominance								1

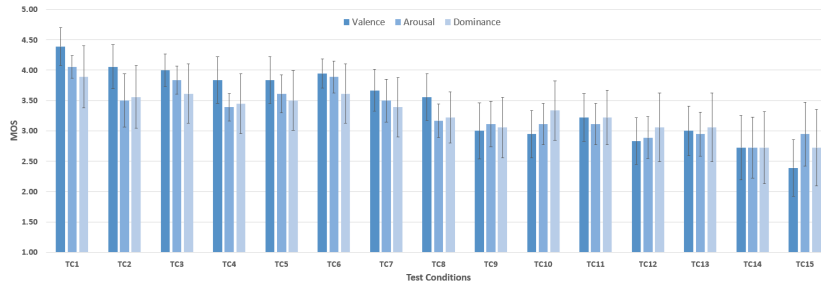


Figure 2.7: MOS with 95% CI for emotions (valence, arousal, and dominance).

2.2.2.2 Emotional Dimensions and Communication Feasibility Analysis-Q3

The emotional response analysis, visualized in Figure 2.7, examines valence, arousal, and dominance across all test conditions. The valence dimension, measuring pleasant-to-unpleasant emotional states, maintained a mean value of 3.43, indicating predominantly positive affect throughout the experiment, even during periods of quality degradation. This finding aligns with observations reported in [47]. Although valence MOS showed inverse correlation with network impairment severity, it consistently exceeded 2.5, dropping below this threshold only under maximum degradation (TC15). Statistical analysis reveals strong correlations between valence and quality metrics (PCC values of 0.980, 0.986, and 0.961 for overall, audio, and video QoE respectively). Additionally, valence demonstrated strong correlations with communication ease (0.984) and game impact (-0.744). The arousal and dominance dimensions exhibited distinct behavioral patterns compared to valence. While arousal quantifies emotional activation levels and dominance reflects perceived situational control, both metrics displayed sinusoidal decay patterns correlating with increased network impairments. This distinctive response pattern resulted in reduced correlations with QoE metrics, particularly for arousal, which demonstrated the lowest correlation coefficients among emotional dimensions.

2.2.3 Outcomes and Discussion

This research has comprehensively investigated the complex interplay between network impairments and user experience in WebRTC-based interactive audiovisual communications. Through rigorous experimental analysis, we examined how delay, jitter, and packet loss affect multimedia quality, emotional dimensions, and overall communication feasibility in real-time conversations. Our findings reveal significant asymmetry in the impact of network impairments across different media components. The experimental results demonstrate that audio QoE consistently achieved superior performance compared to video QoE and overall QoE metrics. This observation carries important implications for the design and optimization of WebRTC applications, suggesting that prioritizing audio quality maintenance during network degradation could help preserve essential communication functionality. A particularly noteworthy discovery emerged regarding the relationship between individual media components and overall quality perception. The arithmetic mean of audio and video QoE scores proved to be a reliable predictor of overall QoE, establishing a practical heuristic for quality estimation in WebRTC systems. This finding offers valuable insights for developing real-time quality monitoring and adaptation mechanisms. The analysis of communication ease revealed a direct correlation between network impairment severity and conversational difficulty. As network conditions deteriorated, participants experienced progressively increasing challenges in maintaining effective communication, though this relationship was not strictly linear. The granular examination of these effects provides crucial understanding for developing more resilient communication systems. In the domain of emotional response, our investigation yielded an unexpected finding regarding user affect under varying network conditions. The valence measurements indicated that participants maintained predominantly pleasant emotional states throughout the experimental sessions, even during periods of significant quality degradation. This resilience in emotional response suggests that users may possess greater tolerance for technical impairments than previously assumed, particularly in task-oriented communication scenarios. These findings collectively advance our understanding of human-centric quality assessment in real-time communication systems. The research establishes a foundation for future work in developing more sophisticated QoE models that consider both technical parameters and human factors in integrated frameworks. Further investigation is warranted to explore potential intrinsic biases in our results, particularly those that might arise from the gaming-based conversational task methodology or the pre-existing familiarity among test participants. The comprehensive insights derived from this study contribute to both theoretical understanding and practical application in the field of real-time communication systems, while simultaneously highlighting important directions for future research in this rapidly evolving domain. Therefore, as a next step we delve into the application-layer data to estimate the QoE in QoE-2 [11].

2.3 Based on Application-layer Data Quality Prediction

The successful deployment of WebRTC-based applications requires appropriate QoE-aware service management approaches to ensure service acceptability from the user’s perspective. Recent research has demonstrated that monitoring and analyzing application-level data can provide valuable insights for estimating the user’s QoE [56–58]. This section presents our investigation into leveraging WebRTC session parameters for QoE estimation through statistical analysis and machine learning approaches with a contribution of the QoE-2 [11].

The research follows a systematic approach to establish robust relationships between WebRTC session parameters and users’ perceived quality. The WebRTC technology enables seamless integration of real-time voice and video communications directly within web browsers, facilitating data packet transmission through peer connections implemented via Real-time Transport Protocol (RTP) over User Datagram Protocol (UDP) protocol. The present investigation focuses on two-party audiovisual conversations, where the communication architecture generates distinct media tracks for each participating endpoint. Specifically, each party maintains four concurrent tracks: two audio tracks and two video tracks, with dedicated channels for both transmission and reception of media content. To enable comprehensive analysis of session performance, the Google Chrome browser provides the *webrtc-internals* diagnostic tool¹, which facilitates real-time monitoring and collection of session-related parameters throughout the communication duration. These parameters encompass multiple categories of data: RTP Stream metrics for both inbound and outbound traffic, including packet transmission statistics, byte counts, and loss metrics; Media Stream characteristics covering both audio and video components, such as audio levels and frame statistics; and Channel-specific information detailing byte transfer rates and protocol utilization.

Given this rich set of available metrics, our research pursued three primary objectives:

- Collection and analysis of application-layer data through the *webrtc-internals* tool during two-party WebRTC audiovisual sessions under controlled network impairment conditions which is collected in the Section 2.2.
- Statistical preprocessing of raw WebRTC session parameters using comprehensive metrics including maximum, minimum, mean, median, standard deviation, variance, and sum.
- Statistical correlation analysis using Analysis of Variance (ANOVA) and PCC to identify parameters significantly related to measured QoE.

¹<chrome://webrtcinternals/>

- Development and validation of machine learning regression models for QoE estimation.

The remainder of this section is structured as follows to present a comprehensive analysis of application-layer data for QoE estimation in WebRTC systems.

Section 2.3.1 elaborates on the statistical analysis performed on the collected WebRTC session data, encompassing both our preprocessing methodology and feature selection approach. Our proposed machine learning-based QoE estimation models are presented in Section 2.3.2, along with a comprehensive discussion of experimental results and performance evaluation. The section 2.3.2 builds upon the statistical foundations established in Section 2.3.1 to develop robust predictive frameworks. Finally, Section 2.3.3 synthesizes our findings and presents concluding remarks, contextualizing our contributions within the broader domain of QoE assessment for real-time communication systems.

2.3.1 Statistical Analysis

The statistical analysis process involved systematic examination of WebRTC session parameters collected during controlled experimental sessions, as illustrated in Fig. 2.8. For each test condition TC_i (with $i \in 1, 2, \dots, 15$), the *webrtc-internals* tool gathered 116 distinct WebRTC session parameters ($p_{i,j,k}$, with $j \in 1, 2, \dots, 116$) from each test participant (u_k , with $k \in 1, 2, \dots, 18$). These parameters were stored as variable-length arrays, with lengths determined by the parameter data characteristics, encompassing information related to audio sources, video sources, and data channel states. The computation of statistical metrics for the WebRTC session parameters followed the procedure detailed in Algorithm 1. Initially, for each parameter collected under identical test conditions $p_{i,j,k}$, we aggregated the parameter data from all 18 test participants. Subsequently, we calculated seven distinct statistical metrics on the aggregated data ($p_{i,j}$): maximum, minimum, mean, median, standard deviation, variance, and sum. These calculations were performed comprehensively across all parameters and test conditions. The processing phase culminated in the generation of seven 15×116 matrices, with each matrix corresponding to a specific statistical metric, as depicted in the lower left section of Fig. 2.8. Throughout the subsequent analysis, we refer to these statistical metrics computed on the WebRTC session parameters as *parameter stats*. Following the initial processing, we proceeded to identify the most significant WebRTC session parameters in relation to the MOS. This identification process employed two statistical methods: the PCC and ANOVA. These analyses were conducted between the parameter stats and the MOS with the objective of establishing potential correlations between the distribution patterns of parameter stats and the mean QoE feedback obtained from test participants.

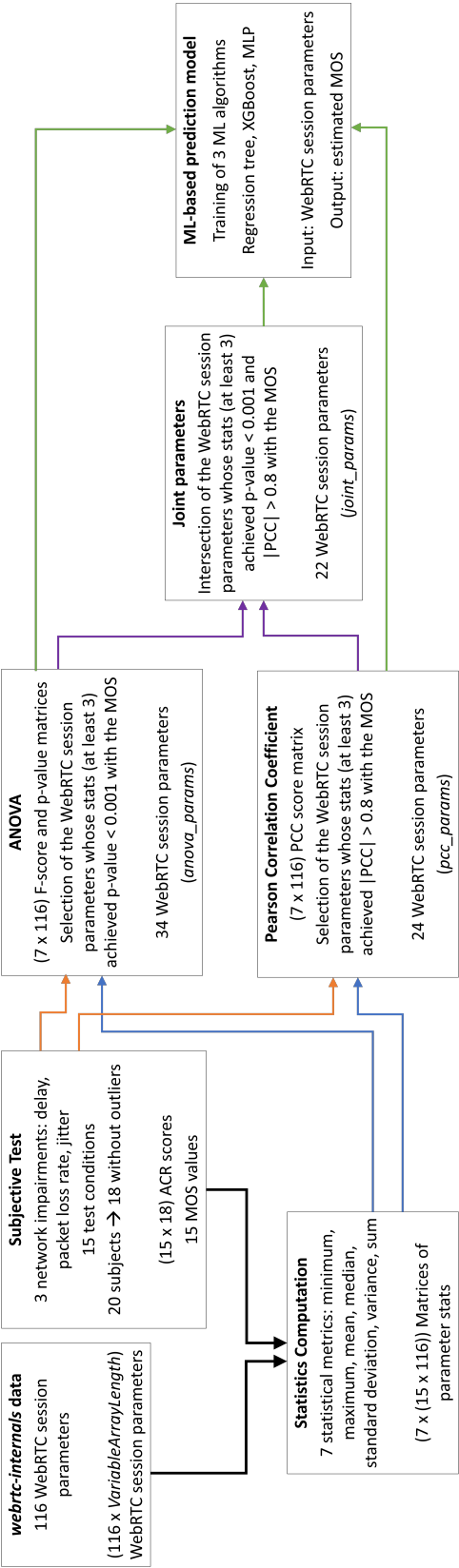


Figure 2.8: Overview of the statistical analysis process for WebRTC session parameters.

Algorithm 1 Calculation of statistical metrics for the WebRTC session parameters.

```

for all  $TC_i$  do                                      $\triangleright i \in \{1, 2, \dots, 15\}$ 
  for all  $p_{i,j}$  do                                      $\triangleright j \in \{1, 2, \dots, 116\}$ 
     $p_{i,j} = \bigcup_{k=1}^{18} p_{i,j,k}$ 
     $p_{i,j}^{min} = \min(p_{i,j})$ 
     $p_{i,j}^{max} = \max(p_{i,j})$ 
     $p_{i,j}^{mean} = \text{mean}(p_{i,j})$ 
     $p_{i,j}^{med} = \text{median}(p_{i,j})$ 
     $p_{i,j}^{stdev} = \text{standardDeviation}(p_{i,j})$ 
     $p_{i,j}^{var} = \text{variance}(p_{i,j})$ 
     $p_{i,j}^{sum} = \text{sum}(p_{i,j})$ 
  end for
end for

```

2.3.1.1 Pearson Correlation Coefficient

The Pearson Correlation Coefficient (PCC) analysis was employed to establish quantitative relationships between measured QoE values and WebRTC session parameters. The PCC provides a standardized measure of the linear correlation between two variables, producing values in the interval $[-1, 1]$. This metric effectively normalizes the covariance between the variables, enabling meaningful comparisons across different parameter scales and units. In our analytical framework, correlation strength was categorized according to established thresholds: values exceeding 0.8 indicated strong positive correlation, while values below -0.8 signified strong negative correlation. For each WebRTC session parameter, we computed a comprehensive set of seven PCC values, corresponding to each derived statistical metric (maximum, minimum, mean, median, standard deviation, variance, and sum). This multifaceted approach enabled us to capture different aspects of the relationship between session parameters and perceived quality. To establish statistical significance, we implemented a rigorous selection criterion: a WebRTC session parameter was classified as correlated with the MOS if at least three of its statistical metrics achieved $|PCC| > 0.8$. For illustration, consider the analysis of the parameter *RTCInbound-RTPAudioStreamInsertedSamplesForDeceleration*, which yielded correlation coefficients of -0.885 for mean, 0.153 for minimum, -0.613 for maximum, -0.860 for median, -0.834 for standard deviation, -0.816 for variance, and -0.900 for sum. With five statistical metrics exceeding our correlation threshold, this parameter demonstrated significant correlation with user-perceived quality. Through systematic application of these analytical criteria across all session parameters, we identified a total of 24 WebRTC parameters demonstrating significant correlation with MOS values. This set of parameters, denoted as *pcc_params*, formed the foundation for subsequent development of our QoE estimation models. The identification of these parameters represents a critical step in establishing robust, data-driven relationships between

Table 2.4: Test conditions (TCs) and MOS.

TC	Delay (ms)	Jitter (ms)	PLR (%)	MOS
1	0	0	0	3.89
2	500	0	0	3.67
3	1000	0	0	3.72
4	500	500	0	3.44
5	1000	500	0	3.44
6	0	0	15	3.22
7	500	0	15	3.06
8	1000	0	15	2.72
9	500	500	15	2.33
10	1000	500	15	2.44
11	0	0	30	2.56
12	500	0	30	2.11
13	1000	0	30	2.11
14	500	500	30	1.89
15	1000	500	30	1.44

objectively measurable session characteristics and subjective quality assessments. The comprehensive correlation analysis enabled us to distill the most relevant parameters from the extensive set of available WebRTC session metrics, providing a statistically sound basis for our machine learning approaches to QoE estimation. This methodical reduction in dimensionality, guided by rigorous statistical criteria, enhanced both the efficiency and interpretability of our subsequent modeling efforts.

2.3.1.2 Analysis of Variance

ANOVA serves as a statistical significance test designed to evaluate differences between multiple population means through variance analysis. In our research context, ANOVA was employed to test the null hypothesis that all population means are equivalent. When this null hypothesis is rejected, it indicates that at least one population mean demonstrates statistical difference from at least one other mean within the analyzed groups. For our analytical framework, we executed ANOVA computations between the collected MOS values and the derived parameter statistics, establishing a significance threshold with $p\text{-value} < 0.001$. This rigorous threshold was selected to ensure high statistical confidence in our findings. In determining significance criteria, we classified WebRTC session parameters as *significant for the MOS* when at least three of their associated statistics demonstrated $p\text{-values}$ below this threshold ($p\text{-value} < 0.001$). Such rejection of the null hypothesis indicates that significant variations in the MOS correspond to statistically significant changes in the parameter statistics' values. To illustrate our analytical approach, we exam-

ined the session parameter *RTCInboundRTPAudioStreamInsertedSamplesForDeceleration*. The p-value analysis for this parameter yielded the following distribution: the mean metric achieved p-value < 0.001 , the minimum metric showed p-value < 0.001 , the maximum metric registered p-value $= 0.0149$, the median calculation produced p-value < 0.001 , the standard deviation analysis indicated p-value < 0.001 , the variance computation demonstrated p-value < 0.001 , and the sum statistic revealed p-value < 0.001 . In this case, six out of seven statistical metrics satisfied our significance threshold, substantially exceeding our minimum requirement of three metrics. Consequently, this parameter was classified as significant for MOS prediction. Through comprehensive application of this ANOVA-based methodology, we identified a total of 34 WebRTC session parameters demonstrating statistical significance for MOS prediction. These parameters, henceforth referred to as *anova_params*, formed a crucial subset of our analytical framework for subsequent machine learning model development.

2.3.2 QoE estimation models

This section presents our comprehensive development and evaluation of machine learning-based QoE estimation models. Our analytical framework utilized WebRTC session parameters collected during 15 test sessions as the feature dataset, with the corresponding 15 MOS values presented in Table 2.4 serving as the target outputs for estimation. To thoroughly investigate the capabilities of PCC and ANOVA in identifying optimal WebRTC parameters for QoE estimation, we developed and evaluated three distinct regression models. Each model was trained using three different feature datasets, resulting in nine trained models overall. The feature datasets were constructed as follows:

- *pcc_params*: the 24 WebRTC session parameters correlated with the MOS identified in the Section 2.3.1;
- *anova_params*: the 34 WebRTC session parameters significant for the MOS identified in the Section 2.3.1;
- *joint_params*: the 22 WebRTC session parameters in common between *pcc_params* and *anova_params*. *joint_params* include the same parameters of *pcc_params* except for 2 parameters²³.

2.3.2.1 Machine Learning Algorithms

Our investigation employed three distinct regression models, each selected for their unique characteristics and capabilities:

²*RTCMediaStreamTrack.sender.hugeFramesSent*

³*RTCRemoteInboundRtpAudioStream-packetsLost*

Table 2.5: The hyperparameters for the three ML methods; optimal values are highlighted in bold. MLP used the ADAM optimizer.

Regression tree		XGBoost			MLP		
Max dep.	Num. leafs	Num. est.	Max dep.	Lrn. rate	Num. lay.	Neur.	Lrn. rate
11	500	700	7	10^{-1}	1	8	10^{-1}
12	600	800	8	10^{-2}	2	16	10^{-2}
13	700	900	9	10^{-3}	3	32	10^{-3}
14	800	1000	10	10^{-4}	4	64	10^{-4}
15	900	1100	11	10^{-5}	5	128	10^{-5}

- Regression tree: A decision tree classifier adapted for continuous variable prediction, where leaf nodes represent class labels and branches encode feature conjunctions leading to those labels.
- eXtreme Gradient Boosting (XGBoost): An implementation of gradient boosting that builds ensemble models from decision trees. This approach incrementally adds trees to correct prediction errors from previous models, utilizing arbitrary differentiable loss functions and gradient descent optimization.
- Multi-Layer Perceptron (MLP): A feedforward artificial neural network architecture implementing regression through backpropagation training. The output layer operates without an activation function (effectively using the identity function), employing square error as the loss function to generate continuous value predictions.

2.3.2.2 Implementation and Optimization Framework

The implementation of our machine learning models was accomplished using Python, with particular attention to standardization and optimization protocols. For feature standardization, we employed the *MinMaxScaler* function from the *scikit-learn* library, ensuring all features were scaled to a comparable range, thereby facilitating more effective model training. Our optimization methodology incorporated a comprehensive hyperparameter tuning process, designed to maximize model performance as measured by Root Mean Square Error (RMSE) (targeting minimal values approaching 0) and coefficient of determination (R^2 , targeting maximal values approaching 1). The optimization framework utilized two complementary approaches from the *scikit-learn* library: *GridSearchCV*, which conducts exhaustive search across specified parameter spaces, and *RandomizedSearchCV*, which implements stochastic parameter selection for improved computational efficiency. To ensure robust model evaluation, we implemented k-fold cross-validation with $k = 5$,

Table 2.6: The 24 *pcc_params*. Note: Dec. is the abbreviation for deceleration.

WebRTC session feature
<i>RTCIceCandidatePair_bytesReceived</i>
<i>RTCInboundRTPAudioStream_concealedSamples</i>
<i>RTCInboundRTPAudioStream_concealmentEvents</i>
<i>RTCInboundRTPAudioStream_headerBytesReceived</i>
<i>RTCInboundRTPAudioStream_insertedSamplesForDec.</i>
<i>RTCInboundRTPAudioStream_jitterBufferEmittedCount</i>
<i>RTCInboundRTPAudioStream_packetsLost</i>
<i>RTCInboundRTPVideo_framesDecoded</i>
<i>RTCInboundRTPVideo_framesDecoded_s</i>
<i>RTCInboundRTPVideo_framesPerSecond</i>
<i>RTCInboundRTPVideo_framesReceived</i>
<i>RTCInboundRTPVideo_headerBytesReceived</i>
<i>RTCInboundRTPVideo_keyFramesDecoded</i>
<i>RTCInboundRTPVideo_pliCount</i>
<i>RTCInboundRTPVideo_totalSquaredInterFrameDelay</i>
<i>RTCMediaStreamTrack_receiver_concealedSamples</i>
<i>RTCMediaStreamTrack_receiver_concealmentEvents</i>
<i>RTCMediaStreamTrack_receiver_insertedSamplesForDec.</i>
<i>RTCMediaStreamTrack_sender_hugeFramesSent</i>
<i>RTCOutboundRTPVideoStream_hugeFramesSent</i>
<i>RTCOutboundRTPVideoStream_keyFramesEncoded</i>
<i>RTCOutboundRTPVideoStream_pliCount</i>
<i>RTCOutboundRTPVideoStream_qpSum_framesEncoded</i>
<i>RTCRemoteInboundRtpAudioStream_packetsLost</i>

partitioning our dataset into 80% training and 20% validation segments. This rigorous validation approach enabled us to develop more reliable model performance estimates by evaluating multiple model configurations. The detailed hyperparameter configurations and their optimal values are presented in Table 2.5.

2.3.2.3 Experimental Results and Analysis

The performance metrics of our three regression methods, trained using the previously described feature datasets and optimized hyperparameters (Table 2.5), are illustrated in Fig. 2.9 and Fig. 2.10. These figures present the mean scores and standard deviations (depicted through error bars) computed across our 5-fold cross-validation procedure. Our analysis revealed that all regression methods achieved optimal performance metrics - both in terms of maximizing R^2 and minimizing RMSE - when trained using the *pcc_params* dataset. This finding strongly suggests that our PCC-based statistical analysis successfully identified the most effective set of WebRTC session parameters for QoE estimation, as detailed in Table 2.6. In-

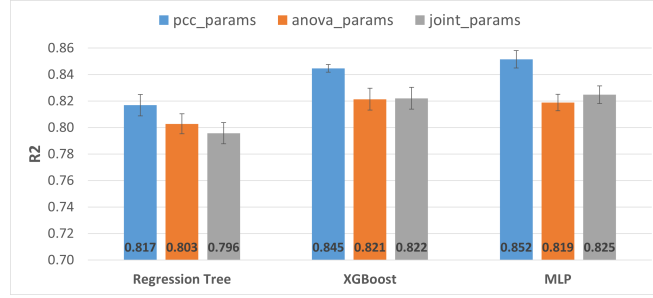


Figure 2.9: Coefficient of determination R^2 results achieved by the 3 regression models when trained on the 3 features datasets.

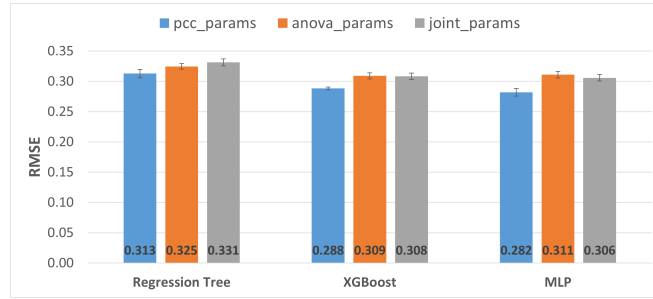


Figure 2.10: RMSE results achieved by the 3 regression models when trained on the 3 features datasets.

terestingly, the inclusion of additional parameters identified through ANOVA (approximately 10 additional features) did not enhance model performance. Instead, these additional features appeared to introduce noise into the system, resulting in degraded estimation performance across all regression models. Furthermore, the intersection of PCC and ANOVA parameters *joint_params* resulted in the loss of two parameters from the original PCC analysis, leading to further performance deterioration. The MLP architecture demonstrated superior performance metrics when trained with both *pcc_params* and *joint_params* datasets, followed by XGBoost and Regression Tree models respectively. However, when training with *anova_params*, XGBoost achieved the highest R^2 and lowest RMSE values. The superior performance of XGBoost compared to the Regression Tree can be attributed to its ensemble learning approach, which combines predictions from multiple decision tree models rather than relying on a single tree structure. This ensemble methodology typically yields enhanced prediction capabilities by leveraging the collective wisdom of multiple models trained on the same dataset.

2.3.2.4 Comparative Analysis with State-of-the-Art Approaches

To contextualize our results within the broader research landscape, we conducted comparative analyses against existing state-of-the-art models. The study by [58]

represents the only comparable work in literature proposing a regression model based on *webrtc-internals* session parameters. Their Multiple Linear Regression approach, utilizing four parameters, achieved an R^2 value of 0.732, which was exceeded by all our proposed models. Furthermore, we evaluated our approach against the DQX model with multiple variables [48], computing estimated MOS values for our 15 test conditions. The resulting RMSE of 0.672 was approximately double that achieved by our proposed models, further validating the effectiveness of our approach.

2.3.3 Outcomes and Discussion

Our research has demonstrated the effectiveness of a comprehensive statistical analysis process incorporating statistical metrics, PCC, and ANOVA for identifying optimal application-level WebRTC session parameters in QoE modeling. The comparative performance of our three regression methods in different feature datasets underscores the critical importance of proper statistical analysis in developing ML-based estimators. The superior performance achieved when training models with PCC-identified parameters, particularly the MLP architecture’s state-of-the-art results ($R^2 = 0.852$, RMSE = 0.282), the next step in the Section 2.4 with a contribution of the study QoE-3 [12] involves QoE estimation based on facial characteristics. This approach aims to leverage the facial expressions of users as a rich source of information, further improving the precision and applicability of QoE prediction models in diverse multimedia contexts.

2.4 Facial Feature-based Quality Prediction

The proliferation of online videoconferencing applications has led to their dominance in consumer Internet traffic, necessitating robust methods for assessing and ensuring service quality. While traditional QoE assessment relies predominantly on subjective quality measurements through direct user feedback, this methodology presents significant limitations for real-time service management systems. This limitation has motivated our investigation into alternative approaches for QoE estimation.

This section presents our novel methodology for estimating QoE in WebRTC-based video communications through automated analysis of users' facial expressions with a contribution of the thesis QoE-3 [12]. Our approach leverages the natural availability of facial data during video calls to develop non-intrusive quality assessment methods. The research comprises three principal components:

- First, we conducted comprehensive subjective quality assessments examining two-party audiovisual conversations over WebRTC, systematically introducing network impairments to evaluate their impact on perceived quality. This experimental framework provided essential ground-truth data for subsequent analysis as explained in the previous section 2.2.
- Second, we developed an innovative feature extraction methodology to capture and quantify facial expressions from participating users during video sessions. These extracted features formed the foundation for our ML approaches to QoE estimation.
- Third, we implemented and evaluated two distinct computational approaches for QoE prediction: a traditional ML algorithm based on SVM classification, and a FCN. Our empirical evaluation demonstrated promising results, with the SVM model achieving a mean estimation accuracy of 0.78 and the FCN reaching 0.70.

The subsequent sections provide detailed examination of these components: Section 2.4.1 presents comprehensive documentation of our ML approach in 2.4.1.2 and FCN-based approach in 2.4.1.3, Section 2.4.2 offers thorough analysis of model performance and comparative evaluation, and Section 2.4.3 synthesizes our findings and discusses their implications for future research.

2.4.1 QoE prediction models

This section outlines the preprocessing of facial features, as well as the ML-based and FCN-based approaches utilized to develop QoE estimation models from facial features.

2.4.1.1 Statistical Analysis of Facial Expression Features

During the experimental phase, we conducted systematic recording and analysis of participants' facial expressions throughout their audio-visual conversations as explained in 2.2. The data collection yielded 270 distinct video recordings, each with a duration of two minutes, captured at full HD resolution (1920x1080) with a frame rate of 30 frames per second. For feature extraction, we employed the OpenFace toolkit [86–88], a specialized framework for facial behavior analysis. The extraction process focused on two primary categories of features: facial expressions and gaze direction characteristics. The facial expression analysis was conducted using the Facial Action Coding System (FACS)⁴ [60], which defines 35 distinct AUs. These AUs comprise two subcategories: 18 AU_c measurements detecting specific muscle activations, and 17 AU_r measurements quantifying muscle activation intensity on a scale from 1 to 5. For Gaze Direction (GD) analysis, the system extracted 6 coordinate arrays representing the directional focus of participants' gaze. The total feature set thus comprised 41 distinct measurements (35 AUs and 6 GD features) per frame, with up to 3600 values extracted per video recording, corresponding to the frame-by-frame analysis. To ensure data quality, we implemented strict confidence thresholds in the OpenFace configuration, setting the face tracking confidence parameter to 98%. This threshold ensured that feature extraction occurred only for frames where facial detection met this high confidence criterion. Consequently, the total number of extracted feature values varied across recordings, dependent on the OpenFace toolkit's ability to maintain reliable face tracking for each participant. To address the high dimensionality of the extracted data while preserving meaningful information, we developed three distinct statistical approaches (S1, S2, and S3) for data reduction:

- S1. We computed the frequency of activation F_{AU_c} for each AU_c , the intensity of activation I_{AU_r} for each AU_r , and the variance of the GD V_{GD} for each gaze feature, following the equations provided in [7].
- S2. We calculated the mean value μ of each of the 41 facial expression features over the (up to) 3600 values extracted for each test.
- S3. We derived the standard deviation σ of each of the 41 facial expression features over the (up to) 3600 values extracted for each test.

This statistical preprocessing phase was crucial for reducing the dimensionality of our feature space while maintaining the discriminative power of the extracted facial and gaze features for subsequent QoE analysis.

We analyzed a comprehensive dataset comprising 11070 features' statistics (15 TCs $\times$ 18 participants $\times$ 41 features) across statistical measures S_1 , S_2 , and S_3 . For each statistical measure, we conducted one-way ANOVA between the features'

⁴<https://www.cs.cmu.edu/~face/facs.htm>

Table 2.7: ANOVA results: features' statistics that achieved a p-value < 0.05 for S_1 , S_2 , and S_3 .

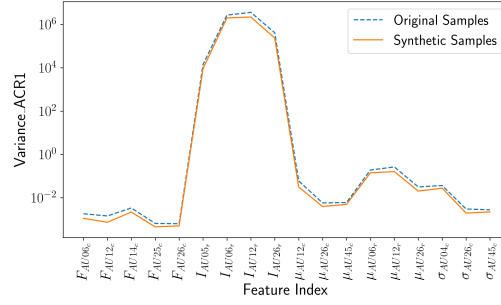
S1	S2						S3		
F_{AU06_c} , F_{AU12_c} , F_{AU14_c} , μ_{AU12_c} , μ_{AU26_c} , μ_{AU45_c} , σ_{AU04_c} , σ_{AU26_c} , σ_{AU45_c}									
F_{AU25_c} , F_{AU26_c} , I_{AU05_r} , μ_{AU06_r} , μ_{AU12_r} , μ_{AU26_r}									
I_{AU06_r} , I_{AU12_r} , I_{AU26_r}									

statistics computed across different TCs and their corresponding ACR scores. Table 2.7 presents the features' statistics that demonstrated statistical significance with p-value < 0.05 across S_1 , S_2 , and S_3 . The analysis revealed that S_1 , S_2 , and S_3 contained 9, 6, and 3 features' statistics respectively that exhibited statistically significant differences in means across different ACR scores, establishing their relevance for QoE estimation model development. A notable pattern emerged in the statistical analysis: S_1 and S_2 statistics encompassed both Action Unit (AU) activation (AU_c) and AU intensity (AU_r), whereas S_3 statistics demonstrated significance exclusively for AU intensity. This analysis identified specific facial muscle movements correlating with QoE perception. Particularly, AU26 (jaw drop) manifested significance across all statistics, while AU06 (cheek raiser) and AU12 (lip corner puller) showed significance in both S_1 and S_2 statistics. Additionally, AU45 (blink) demonstrated significance in both S_2 and S_3 statistics. Notably, eye gaze features showed no significant correlation, indicating their limited utility for QoE estimation in this context.

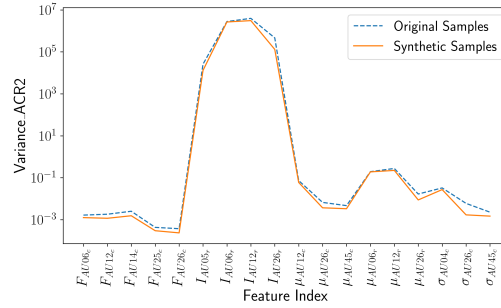
2.4.1.2 ML-based QoE models based on the facial features

Prior to training the ML-based models with the identified facial features' statistics, we implemented data augmentation to address the dataset's class imbalance. The distribution of ACR scores exhibited under representation of extreme scores compared to middle scores, as detailed in Table 2.8. To address this imbalance, we employed the Adaptive Synthetic Algorithm (ADASYN) algorithm for synthetic sample generation and class over-sampling. ADASYN enhances class balance through adaptive generation of synthetic instances from minority classes, utilizing linear weighted interpolation between existing minority class examples to mitigate bias from imbalanced data distribution [89]. Table 2.8 demonstrates the improved class distribution post-augmentation. Figure 2.11 presents a comparative analysis of variance between original facial features' statistics and synthetic features' statistics generated via ADASYN for ACR scores 1, 2, 4, and 5, confirming that synthetic samples maintain consistent distribution characteristics with the original samples.

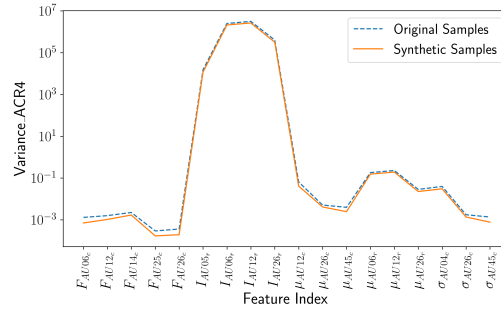
For QoE estimation model development, we structured the features into four distinct subsets based on the significant features' statistics identified in Section 2.4.1.1: In particular, we considered the following subsets of features (see Table 2.7) to train



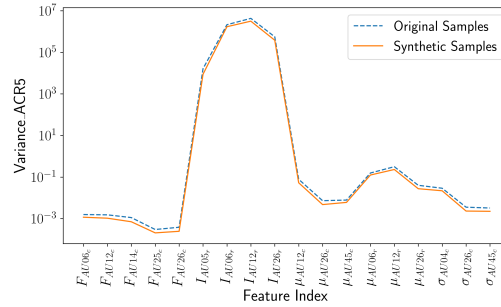
(a) ACR1.



(b) ACR2.



(c) ACR4.



(d) ACR5.

Figure 2.11: Variance of the original and synthetic facial expression features' statistics for the ACR scores 1, 2, 4, and 5. Note that a logarithmic scale was used for the y-axis.

Table 2.8: Number of ACR scores (and corresponding facial features) before and after data augmentation.

ACR	Collected samples	Augmented samples
1	37	105
2	65	121
3	103	103
4	44	101
5	21	99
Total	270	529

the classifiers:

- Subset1: The 9 S_1 feature’s statistics.
- Subset2: The 6 S_2 feature’s statistics.
- Subset3: The 3 S_3 feature’s statistics.
- Subset4: The 18 $S_1 + S_2 + S_3$ feature’s statistics.

The classifier development employed a 70%/30% training/validation split with 5-fold cross-validation. Among various classifier architectures evaluated (including decision trees, naive Bayes, ensembles, discriminant analysis), k-nearest neighbors (k-NN) and SVM demonstrated superior QoE estimation performance. The optimal classifier configurations were determined as follows:

- Subset1 and Subset2: the k-NN with $n = 1$, utilizing the Euclidean distance metric and uniform weighting distance.
- Subset3: the k-NN with $n = 13$, utilizing the Euclidean distance metric and the inverse squared weighting distance.
- Subset4: the SVM with the linear kernel function and a kernel scale of 1.

2.4.1.3 FCN-based QoE model

In addition to the ML algorithms, we implemented a FCN approach for QoE estimation. While a traditional CNN architecture integrates both convolutional and fully connected layers, this structure presents inherent limitations due to its requirement for uniform input matrix dimensions. As previously detailed in the Section 2.4.1.1, the variable number of features extracted during different web calls results in dataset input matrices of heterogeneous sizes, presenting a significant challenge for conventional CNN training methodologies. The implementation of a CNN for analyzing AU and gaze matrices would necessitate either matrix cropping or zero-padding

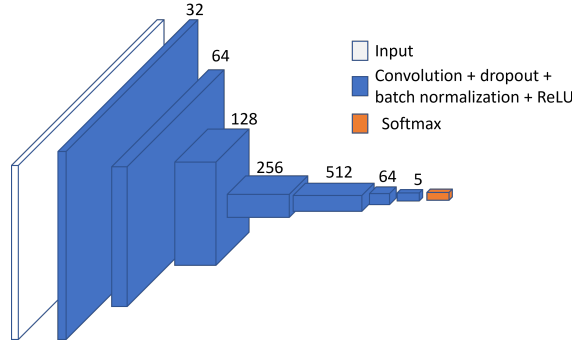


Figure 2.12: FCN architecture. On the top of each box is shown the number of filters used in the convolution step.

to achieve uniform dimensionality. To address these architectural constraints, we developed an FCN-based QoE estimator. The FCN architecture eliminates fully connected layers, replacing them with convolutional layers of 1×1 dimensionality. This structural modification resolves the uniform matrix size requirement inherent to fully connected layers. A fundamental distinction of the FCN methodology lies in its capacity for full dataset training without prerequisite facial feature preprocessing. The FCN filters autonomously process facial input matrices to identify salient information pertinent to QoE classification. This approach contrasts with the preprocessing methods outlined in Section 2.3.1, which reduce up to 3600 feature values to single metrics. While such preprocessing facilitates data reduction and significant feature identification, it potentially results in information loss through approximation of feature trends across the two-minute web call duration. The FCN methodology enables utilization of complete feature vectors throughout the web call duration, delegating computational optimization to the neural network architecture. Furthermore, the FCN processes all available features, whereas ML-based models restrict input to features demonstrating statistical significance through ANOVA analysis. Our FCN architecture, based on the design proposed in [90], implements progressively increasing convolutional filters followed by strategic dimensionality reduction to emulate fully connected layer functionality. Through extensive architectural exploration and parameter optimization, we developed a seven-layer convolutional structure. Each convolutional step incorporates dropout regularization, batch normalization, and ReLU activation functions. The complete architectural framework is illustrated in Fig. 2.12. The model training protocol maintained consistency with previous methodologies, employing a 70%/30% training/validation split and 5-fold cross-validation for robust performance evaluation.

2.4.2 Performance of QoE prediction models

This section presents a detailed analysis of the QoE estimation performance achieved by both the ML-based models and the FCN-based model. The comparative evalua-

tion metrics, including mean accuracy, accuracy, precision, recall, and F1-score for individual ACR scores, are summarized in Table 2.9.

2.4.2.1 ML-based Models Performance Analysis

The experimental results demonstrate that Method 4, which incorporates all features identified by the preceding three methods, achieves superior performance across all evaluation metrics. This comprehensive approach yields the highest mean accuracy and consistently outperforms other methods in terms of accuracy, precision, recall, and F1-score across individual ACR scores. The performance advantage of Method 4 can be attributed to its utilization of a more extensive set of significant features, enabling the ML models to develop more robust classification capabilities. A notable pattern emerges in the classification accuracy across different ACR scores. The model exhibits exceptional performance for extreme classes (ACR 1 and 5) with accuracy approaching 1.0, while maintaining strong performance for intermediate classes (ACR 2 and 4) with accuracy around 0.8. However, the accuracy decreases towards the middle class (ACR 3), reaching its lowest point at 0.35. This trend is consistent across all methods, though with varying degrees of absolute performance. Among the initial three methods, Method 1 demonstrates superior estimation performance, suggesting its preprocessing approach is particularly well-suited for facial feature analysis. Method 2 and Method 3 follow in descending order of performance, with their limitations primarily attributable to their identification of fewer significant features compared to Method 1. Method 3, in particular, exhibits reduced performance due to its reliance on only three significant features. Despite their individual limitations, Methods 2 and 3 contribute valuable complementary information when combined with Method 1’s features. This synergistic effect is evident in Method 4’s enhanced performance, which successfully integrates all identified features from the previous methods. The only notable limitation in Method 4’s performance is its reduced accuracy for the middle class (ACR 3). Nevertheless, the model’s ability to accurately distinguish between satisfied (ACR 4 and 5) and dissatisfied (ACR 1 and 2) users represents a significant achievement, particularly considering its reliance solely on facial expressions without requiring explicit user feedback. An interesting observation regarding algorithm selection emerges from the analysis. While k-NN proves most effective for the first three methods, SVM demonstrates superior performance when applied to Method 4’s comprehensive feature set. This transition likely reflects the SVM’s particular effectiveness in handling larger datasets with complex spatial distributions of features.

2.4.2.2 FCN-based Model Performance Analysis

The FCN-based model achieves a mean accuracy of 0.70, comparable to Method 1 but falling short of Method 4’s 0.78. However, it exhibits a distinctly different pattern in class-specific accuracy distribution. The FCN achieves its highest accuracy

Table 2.9: QoE estimation performance achieved by the ML-based models and the FCN-based model.

Estimation Model	Metric	ACR score				
		1	2	3	4	5
Method 1 k-NN	Mean Acc.	0.70				
	Accuracy	0.80	0.78	0.30	0.75	0.86
	Precision	0.71	0.72	0.49	0.65	0.86
	Recall	0.80	0.78	0.30	0.75	0.86
	F1-Score	0.75	0.75	0.37	0.70	0.86
Method 2 k-NN	Mean Acc.	0.60				
	Accuracy	0.72	0.57	0.20	0.65	0.88
	Precision	0.65	0.56	0.30	0.60	0.8
	Recall	0.72	0.57	0.20	0.65	0.88
	F1-Score	0.68	0.57	0.24	0.62	0.84
Method 3 Weighted k-NN	Mean Acc.	0.42				
	Accuracy	0.47	0.33	0.16	0.38	0.77
	Precision	0.45	0.36	0.17	0.38	0.67
	Recall	0.47	0.33	0.16	0.38	0.77
	F1-Score	0.46	0.34	0.17	0.38	0.72
Method 4 SVM	Mean Acc.	0.78				
	Accuracy	0.97	0.79	0.35	0.81	0.98
	Precision	0.86	0.71	0.62	0.75	0.89
	Recall	0.97	0.79	0.35	0.81	0.98
	F1-Score	0.91	0.75	0.45	0.78	0.93
FCN	Mean Acc.	0.70				
	Accuracy	0.54	0.64	0.82	0.68	0.71
	Precision	0.90	0.52	0.70	0.85	0.90
	Recall	0.54	0.64	0.82	0.68	0.71
	F1-Score	0.70	0.57	0.76	0.76	0.83

(0.82) for the middle class, with performance decreasing towards extreme classes asymmetrically. Specifically, ACR 1 shows the lowest accuracy (0.54), while ACR 5 maintains relatively strong performance (0.71). Intermediate classes ACR 2 and 4 achieve accuracies of 0.64 and 0.68, respectively. The FCN model's training process notably differs from the ML models in its approach to data augmentation. While ML models benefit from ADASYN-augmented preprocessed features, the FCN utilizes the original dataset to avoid potential artifacts in the augmented data that could compromise performance. This methodological difference partly explains the FCN's superior performance on class 3, which dominated the original imbalanced dataset. While the FCN shows promise, its performance is currently constrained by the limited size of the available feature dataset. The architecture's potential for en-

hanced estimation accuracy could be better realized with a larger training dataset, suggesting a direction for future improvement.

2.4.3 Outcomes and Discussion

The research presents novel QoE estimation models leveraging ML algorithms and a FCN trained with facial expression features. To establish ground-truth QoE measurements, a comprehensive subjective assessment was conducted examining audiovisual conversations over a WebRTC-based telemeeting service under controlled network impairment conditions. The methodology for the ML-based models incorporated four distinct preprocessing approaches for the facial expression features prior to model training. To address dataset imbalance concerns, data augmentation techniques were applied to enhance the training dataset distribution. The initial three preprocessing methods identified statistically significant features through systematic analysis of the collected data. Among these approaches, Method 1 demonstrated superior performance with a mean accuracy of 0.70. However, Method 4, which incorporated the collective significant features identified across all three preceding methods, achieved markedly improved results with a mean accuracy of 0.78. This enhanced performance can be attributed to the expanded training dataset encompassing the most salient features identified through the comprehensive analysis of all three methods. In contrast, the FCN implementation utilized the complete facial features dataset without preprocessing. This approach precluded the application of data augmentation techniques, resulting in performance characteristics that reflected the inherent dataset imbalance. While achieving an overall mean accuracy of 0.70, the model exhibited peak performance (0.82 accuracy) for ACR class 3, which represented the most populous category in the dataset. The FCN approach demonstrated limitations primarily attributed to constraints in dataset size. Future research directions will explore the integration of speech features into the estimation models. Speech characteristics provide additional indicators of emotional state, and the incorporation of these supplementary features may enhance the predictive capabilities of both the ML algorithms and particularly the FCN through the expanded feature space. This multimodal approach presents promising opportunities for advancing QoE estimation accuracy as we extend our research in multimodal approach for QoE-4.

2.5 Multimodal Quality Prediction

The use of facial and speech-related features for estimating QoE in multimedia services represents an emerging but underexplored research direction with significant potential. Current approaches have been limited to the analysis of facial expressions or speech characteristics in isolation, with only preliminary experimental validation conducted so far. To advance the state-of-the-art in this domain, this research

with a contribution of the thesis QoE-4 [13] focuses on WebRTC-based videoconferencing applications, which provide a natural environment for capturing both facial expressions and vocal speech characteristics of users during their interactions. By leveraging these naturally occurring behavioral signals, we can develop more robust and comprehensive QoE estimation approaches. The research methodology follows a systematic three-phase approach:

1. The first phase involves extensive statistical analysis to identify the most significant facial (see in 2.4.1.1) and speech-related features for QoE estimation. These features are extracted from participants' audio-visual data collected during carefully designed subjective assessment experiments (see in 2.2.1.3). The statistical significance testing ensures that only the most relevant features are considered for subsequent modeling.
2. In the second phase, we develop individual ML-based QoE estimation models trained separately on the facial (see in 2.4.1.2) and speech datasets. This phase establishes baseline performance metrics and validates the predictive capability of each modality independently. The facial features are extracted using the FACS methodology, while speech features encompass both prosodic and acoustic characteristics.
3. The third phase QoE-4 introduces novel data fusion techniques to combine the facial and speech datasets into a unified representation. Specifically, we employ the Improved Centered Kernel Alignment (ICKA) algorithm to optimally integrate the complementary information provided by both modalities. This multimodal fusion approach demonstrates superior performance, achieving a mean QoE estimation accuracy of 0.93, compared to 0.78 and 0.86 when using only facial or speech features respectively. These results highlight the significant advantages of combining multiple behavioral signals for robust QoE estimation in multimedia communication systems.

The remainder of this section is structured as follows: First, we present the methodology utilized for combining facial and speech features in Section 2.5.1, detailing our approach to feature extraction, preprocessing, and data fusion techniques. Section 2.5.2 discusses QoE estimation models based solely on speech features, analyzing their performance metrics and limitations. Next, Section 2.5.3 presents our integrated approach combining both facial and speech features, describing the novel fusion techniques and examining the resulting performance improvements. Section 2.5.4 provides a comprehensive comparison with state-of-the-art methods, demonstrating the advantages of our multimodal approach. Finally, Section 2.3.3 summarizes the key findings and insights from our facial and speech-based QoE estimation research and the 2.

Table 2.10: Test conditions (TCs) and MOS with 95% CI.

TC	Delay (ms)	Jitter (ms)	PLR (%)	MOS	CI 95%
1	0	0	0	3.89	± 0.404
2	500	0	0	3.67	± 0.511
3	1000	0	0	3.72	± 0.338
4	500	500	0	3.44	± 0.414
5	1000	500	0	3.44	± 0.316
6	0	0	15	3.22	± 0.329
7	500	0	15	3.06	± 0.392
8	1000	0	15	2.72	± 0.207
9	500	500	15	2.33	± 0.344
10	1000	500	15	2.44	± 0.492
11	0	0	30	2.56	± 0.316
12	500	0	30	2.11	± 0.404
13	1000	0	30	2.11	± 0.459
14	500	500	30	1.89	± 0.404
15	1000	500	30	1.44	± 0.352

2.5.1 Methodology

The research methodology encompasses a comprehensive subjective quality assessment protocol wherein participants engage in conversation pairs using a WebRTC-based application platform as explained in Section 2.2.1.3. To systematically evaluate quality perception under varying conditions, the network connection was subjected to controlled impairment through combinations of degradation factors as seen in Table 2.10, specifically engineered to affect audio-visual call quality. Throughout each session, both facial expressions and speech patterns of participants were recorded, generating rich multimodal data streams for subsequent feature extraction and analysis.

Figure 2.13 illustrates the proposed methodological framework. The dashed components indicate the collection of ACR scores—subjective ratings provided by users which are exclusively required during the training phase. These scores serve dual purposes: first, as ground truth data for identifying the most relevant facial and speech features for QoE estimation; and second, as training data for developing ML-based models utilizing these selected significant features.

A comprehensive analysis of speech characteristics was conducted using the OpenSmile software framework to extract 100 distinct psychoacoustic sharpness features from participants' recorded speech. The application of K-means clustering to these features revealed three distinct clusters, effectively categorizing participants into differentiated groups based on voice sharpness characteristics. To reduce the dimensionality of the feature space while preserving essential information, Principal

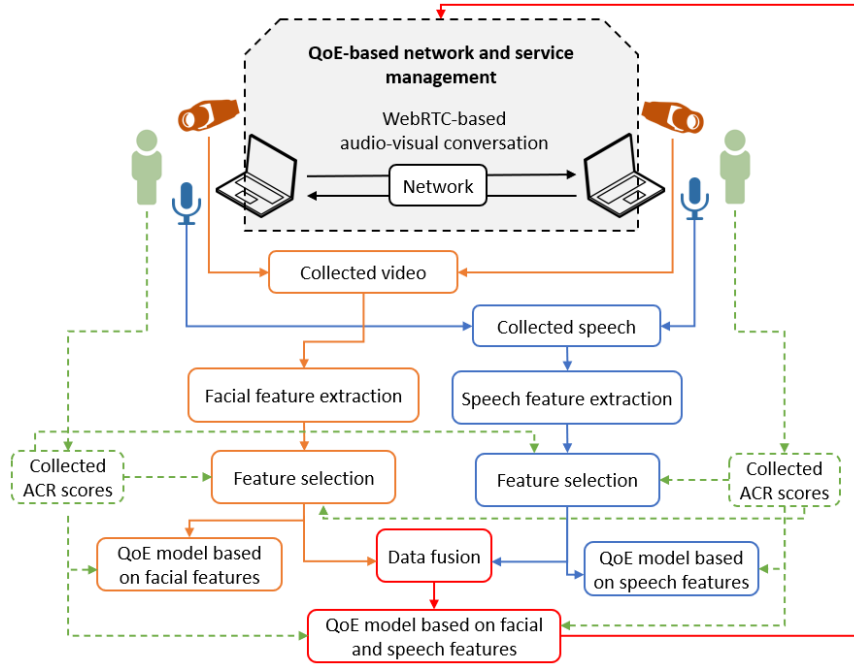


Figure 2.13: The proposed methodology. The dashed lines indicate the collection of ACR scores, which are only needed for the training phase.

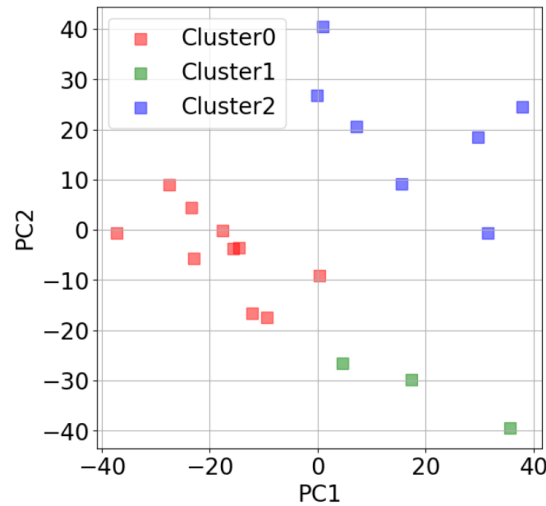


Figure 2.14: The three clusters identified by clustering on the 100 psychoacoustic sharpness features extracted from the recorded speech of the 20 participants. The PCA was used to reduce feature dimensionality to the two principal components (PC1 and PC2).

Component Analysis (PCA) was applied, reducing the 100-dimensional feature space to two principal components (PC1 and PC2), which collectively account for 75% of

the total feature variance. This dimensionality reduction is visualized in Figure 2.14, where the three identified clusters demonstrate clear separation in the reduced feature space. Further statistical validation through ANOVA between the complete feature set and the identified clusters revealed three features of particular significance (p-value < 0.001):

- *pcm_fftMag_psySharpness_sma_de_maxSegLen*
- *pcm_fftMag_psySharpness_sma_de_segLenStddev*
- *pcm_fftMag_psySharpness_sma_de_peakMeanRel*

These features demonstrated robust capability in discriminating between different speech sharpness groups, providing a reliable basis for subsequent analysis.

The developed QoE model has significant operational applications, particularly in network and service management contexts. During operational deployment, the model can provide real-time QoE estimates based on observed facial and speech features, enabling proactive management responses to detected quality degradations. This approach's versatility is enhanced by its compatibility with traditional network- and service quality-related parameters, allowing for comprehensive quality monitoring and management strategies. The integration of this QoE estimation approach with network and service performance indicators enables sophisticated root cause analysis and predictive quality management. This synthesis of behavioral and technical metrics provides a foundation for developing anticipatory approaches to quality issues and optimizing resource allocation within deployed services.

2.5.2 QoE prediction based on the speech features

This section presents our methodology for estimating QoE through speech-based features, encompassing both the statistical analysis framework (see in the Section 2.5.2.1) and the performance evaluation of ML-based QoE models trained on selected speech features (see in the Section 2.5.2.2).

2.5.2.1 Statistical analysis of speech features

As outlined in Section 2.4.3, the experimental protocol incorporated comprehensive audio recordings of participant speech during audio-visual conversations. To ensure robust analysis, we developed three distinct versions of the speech files for comparative assessment:

- **Original Speech (OS):** The OS comprised the unmodified recordings captured during the experimental sessions, preserving both conversational speech and naturally occurring silent intervals when participants assumed listening roles, along with ambient background noise inherent to the recording environment.

Table 2.11: ANOVA results: LLDs that contain functional statistical features with a p-value < 0.01 for the 3 speech files: OS, NRS, and NSS.

LLD	OS	NRS	NSS
audspec_lengthL1norm	7	4	1
audSpec_Rfilt	8	49	10
audspecRasta	-	7	4
F0final	-	2	3
jitterLocal	1	-	9
jitterDDP	-	-	3
logHNR	-	3	2
mfcc_sma	43	20	25
pcm_fftMag	44	16	14
pcm_RMSenergy	3	3	3
pcm_zcr	5	3	-
shimmerLocal	-	-	4
voicingFinalUnclipped	2	4	8
Total	113	111	86

- For the Noise Reduced Speech (NRS), it is implemented a sophisticated non-stationary noise reduction methodology utilizing spectral gating. This processing was executed through the noisereduce⁵ Python algorithm [91, 92], which employs dynamic threshold adjustment in the time domain to attenuate noise components while preserving speech integrity.
- The Non-Silent Speech (NSS) version was derived by algorithmically removing silent intervals from the OS recordings using the librosa⁶ Python library [93], resulting in continuous speech segments.

We employed the OpenSMILE feature extraction toolkit [94, 95] to extract comprehensive speech features from each version. Our analysis focused on low-level descriptor (LLD) contours with functionals based on the ComParE acoustic feature set [96, 97]. The feature extraction process yielded 64 LLDs per speech file, comprising 4 energy-related, 55 spectral, and 6 voicing characteristics. These were further processed to generate 6,373 functional statistical features for each speech version.

The complete dataset encompassed 1,720,710 feature statistics values, derived from 15 test conditions across 18 participants, with 6,373 features per condition. We conducted one-way ANOVA between the speech features (grouped by test conditions) and corresponding ACR scores. Table 2.11 presents the LLDs containing functional statistical features achieving statistical significance ($p < 0.01$) for each

⁵<https://pypi.org/project/noisereduce/>

⁶<https://github.com/librosa/librosa>

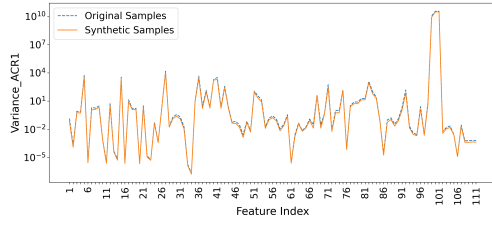
Table 2.12: Number of ACR scores (and corresponding speech features) before and after data augmentation for the 3 speech files: OS, NRS, and NSS.

ACR score	Collected samples	Augmented Samples		
		OS	NRS	NSS
1	37	104	102	107
2	65	107	110	107
3	103	103	103	103
4	44	111	104	105
5	21	104	98	97
Total	270	529	517	519

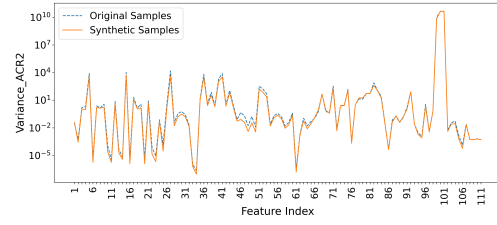
speech version. This significance threshold, while lower than that employed in Section 2.4.1.1, still yielded substantial significant features: 113 for OS, 111 for NRS, and 86 for NSS.

2.5.2.2 ML-based QoE models based on the speech features

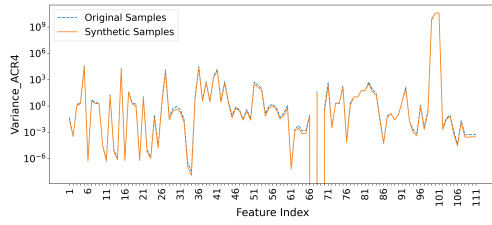
Prior to training our ML-based models with the identified speech feature statistics, we addressed dataset class imbalance through ADASYN-based data augmentation. Table 2.12 presents the ACR score distribution before and after augmentation for each speech version, with variations in augmented samples reflecting the differing numbers of significant features identified per version. Figure 2.15 demonstrates the statistical alignment between original and synthetic feature distributions across ACR scores 1, 2, 4, and 5. We evaluated multiple ML algorithms including SVM, k-NN, decision trees, naive Bayes, ensembles, and discriminant analysis, implementing 70%/30% training/validation partitioning with 5-fold cross-validation. The feature sets comprised 113 significant features for OS, 111 for NRS, and 86 for NSS. The SVM with linear kernel function and unity box constraint level achieved optimal estimation performance across all feature sets. Secondary performance was achieved by k-NN with Euclidean distance metric, followed by SVM with cubic kernel function. Table 2.13 presents comprehensive performance metrics for the optimal classifier configuration. The NRS-trained SVM achieved superior mean accuracy, maintaining minimum recall of 0.70 across all ACR classes, whereas OS and NSS models exhibited lower recall (< 0.60) for class 3. This establishes NRS features as most suitable for QoE estimation. All models demonstrated robust recall (> 0.80) for extreme classes (1, 4, 5), with reduced performance for intermediate classes (2, 3), suggesting enhanced emotional feature salience at quality extremes - a pattern consistent with facial feature-based estimators.



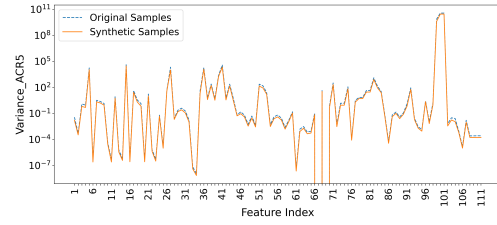
(a) ACR1.



(b) ACR2.



(c) ACR4.



(d) ACR5.

Figure 2.15: Variance of the original and synthetic speech features' statistics for ACR scores 1, 2, 4, and 5. Note that a logarithmic scale was used for the y -axis.

Table 2.13: QoE estimation performance achieved by the ML-based models trained with the speech features.

Speech file / ML model	Performance metric	ACR score				
		1	2	3	4	5
OS SVM	Mean Acc.			0.83		
	Precision	0.85	0.71	0.76	0.80	0.99
	Recall	0.95	0.76	0.52	0.89	0.99
	F1-Score	0.90	0.73	0.62	0.85	0.99
NRS SVM	Mean Acc.			0.86		
	Precision	0.98	0.89	0.61	0.99	0.99
	Recall	0.93	0.70	0.88	0.83	0.99
	F1-Score	0.95	0.78	0.72	0.90	0.99
NSS SVM	Mean Acc.			0.85		
	Precision	0.93	0.79	0.76	0.83	0.93
	Recall	0.93	0.85	0.57	0.91	0.98
	F1-Score	0.93	0.82	0.65	0.87	0.95

2.5.3 Facial & Speech Fusion: Elevating Quality Predictions

Previous section demonstrated that individual modality approaches for QoE estimation achieved promising but limited results, with SVM-based models trained on facial and speech features reaching accuracies of 0.78 and 0.86 respectively. While these results validated the potential of behavioral feature analysis for QoE estimation, they also suggested opportunities for enhancement through multimodal integration. This section presents a comprehensive multimodal fusion framework that leverages complementary information from both facial and speech domains to achieve more robust QoE estimation.

2.5.3.1 Input features and data augmentation

The foundation of our multimodal approach begins with two distinct feature matrices: facial features dataset $\mathbf{FAC}_{(270,18)}$ and speech feature dataset $\mathbf{SP}_{(270,111)}$. The facial feature set comprises 18 carefully selected features identified through statistical analysis (detailed in Section 2.4.1.2), while the speech domain contributes 111 statistically significant features extracted from noise-reduced speech signals (described in Section 2.5.2.1). These features correspond to 270 subjective quality ratings captured through ACR scores, represented in the label matrix $\mathbf{L}_{(270,1)}$. A critical challenge in multimodal fusion involves addressing data imbalance while maintaining feature correspondence on the $\mathbf{FAC}$ and $\mathbf{SP}$ datasets. Rather than independently augmenting each modality, which would risk destroying cross-modal relationships, we developed an integrated augmentation strategy. This approach begins by creating a unified representation through horizontal concatenation on the $\mathbf{FAC}$ and $\mathbf{SP}$

Table 2.14: Number of ACR scores (and corresponding facial and speech features) before and after data augmentation.

ACR	Collected samples	Augmented samples
1	37	108
2	65	108
3	103	103
4	44	106
5	21	96
Total	270	521

datasets as in eq. (2.1) to create the **FAC-SP** single dataset:

$$\mathbf{FAC-SP}_{(270,129)} = [\mathbf{FAC}_{(270,18)} | \mathbf{SP}_{(270,111)}]. \quad (2.1)$$

The concatenated matrix preserves the natural co-occurrence relationships between facial and speech features while providing a foundation for balanced augmentation. We then apply the ADASYN algorithm to generate synthetic samples that maintain these cross-modal relationships on the **FAC-SP** dataset and the corresponding **L** matrix, as in eq. (2.2):

$$[\mathbf{FAC}_{aug}, \mathbf{SP}_{aug}, \mathbf{L}_{aug}] = \text{ADASYN}(\mathbf{FAC-SP}, \mathbf{L}), \quad (2.2)$$

This process yields augmented feature matrices $\mathbf{FAC}_{aug(521,18)}$ and $\mathbf{SP}_{aug(521,111)}$ alongside corresponding labels $\mathbf{L}_{aug(521,1)}$. To ensure consistent feature scaling across modalities, we apply Z-score normalization to both $\mathbf{FAC}_{aug(521,18)}$ and $\mathbf{SP}_{aug(521,111)}$ augmented matrices. Table 2.14 illustrates the effectiveness of this approach in achieving balanced class distributions while preserving the integrity of cross-modal relationships.

2.5.3.2 Data Fusion Architecture

The superior performance of SVM classifiers on individual modalities motivates their adoption within our fusion framework. However, effective multimodal integration requires careful consideration of how different feature types interact and complement each other. We investigate two distinct fusion approaches that offer complementary advantages: dimensionality reduction through PCA and kernel-space integration via ICKA.

Principal Component Analysis Integration: The integration of multimodal features through PCA provides a statistically rigorous framework for capturing the essential variance structure across both facial and speech domains while addressing the curse of dimensionality. This approach leverages PCA’s capacity to identify and

preserve the most informative patterns within high-dimensional data through orthogonal transformation, particularly valuable when combining heterogeneous feature sets. The fundamental principle underlying our PCA-based fusion strategy stems from the observation that while both facial and speech features contain valuable information for QoE estimation, they likely exhibit internal redundancies and cross-modal correlations. Rather than applying fusion directly to the raw feature spaces, we first perform independent dimensionality reduction on each modality to extract their core representational structures:

$$\mathbf{FACaug} \xrightarrow{\text{PCA}} \mathbf{FACaug}^{PCA}, \quad \mathbf{SPAUG} \xrightarrow{\text{PCA}} \mathbf{SPAUG}^{PCA}$$

Through empirical analysis using the elbow method, we determined that the first four principal components capture approximately 95% of the variance in both feature spaces. This finding suggests a remarkably efficient compression of the essential information content, reducing the feature dimensionality while preserving the discriminative power of each modality. The compressed representations are then integrated through horizontal concatenation:

$$\mathbf{FAC-SPAUG}^{PCA} = [\mathbf{FACaug}^{PCA} | \mathbf{SPAUG}^{PCA}]. \quad (2.3)$$

The unified feature representation undergoes SVM classification using a rigorous training protocol with 5-fold cross-validation and a 70%/30% training-validation partition. To ensure optimal model performance, we conducted extensive hyperparameter optimization through GreedSearchCV function⁷ across three critical dimensions:

- Penalty parameter $C \in 0.1, 1, 10, 100, 1000$
- Kernel coefficient $\gamma \in 1, 0.1, 0.01, 0.001, 0.0001$
- Kernel functions: Gaussian, linear, and radial basis

Empirical evaluation revealed optimal performance with $C = 10$, $\gamma = 1$, and the radial basis kernel function and $K = \text{radial basis}$, as documented in Table 2.15.

Improved Centered Kernel Alignment Framework: While PCA provides effective dimensionality reduction, the ICKA [98] framework offers a complementary approach by operating directly in the kernel space, allowing for more nuanced modeling of non-linear relationships between modalities. The ICKA methodology extends traditional kernel alignment by incorporating ideal kernel blocks, thereby optimizing the fusion weights based on the discriminative power of each modality. The theoretical foundation begins with the dual formulation of the SVM optimization problem,

⁷https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

employing the kernel trick to implicitly map features into a high-dimensional space where linear separation becomes feasible. The optimization objective is formulated as [99]:

$$\max_{\alpha} \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j y_i y_j k(x_i x_j), \quad (2.4)$$

subject to positivity and balance constraints:

$$\alpha_i \geq 0, i = 1, \dots, m, \sum_{i=1}^m \alpha_i y_i = 0, \quad (2.5)$$

where $\alpha_i \alpha_j$ are Lagrange multipliers, x is the set of samples, m is the total number of samples in the set of features x , y is the set of labels, and $k(x_i x_j)$ is the kernel trick non-linear mapping function. The core innovation of ICKA lies in its construction of a fusion kernel through weighted combination of modality-specific base kernels, as follows:

$$k(x_i, x_j) = \sum_{h=1}^H d_h k_h(x_i, x_j), \text{ s.t. } d_h \geq 0, \text{ and } \sum_{h=1}^H d_h = 1, \quad (2.6)$$

where d_h is the non-negative weight for the base kernel and H is the total number of base kernels. These weights are computed through a sophisticated distance metric that measures the alignment between base kernels and ideal kernel blocks. d_h is obtained as follows:

$$d_h = \text{dist}(K_h) / \sum_{h=1}^H \text{dist}(K_h). \quad (2.7)$$

with the distance function *dist* incorporating class-specific information which is measured between the base kernel and the ideal kernel, which is defined as follows:

$$\text{dist}(K_h) = \sum_{c=1}^C S(|D_{hc}|) / S(|K_h|) \quad (2.8)$$

where $h = \{1, 2, \dots, H\}$, $c = \{1, 2, \dots, C\}$ identifies the number of classes, S is the sum function of all elements of the matrix D_{hc} , and D is the diagonal block element that identifies the class c and base kernel h . D_{hc} is obtained from the ideal kernel block K_{block} defined as:

$$K_{block}(X, X) = \begin{bmatrix} D_{h1} & & \\ & \ddots & \\ & & D_{hc} \end{bmatrix} \quad (2.9)$$

where $h = \{1, 2, \dots, H\}$ and $c = \{1, 2, \dots, C\}$. Therefore, according to eq. (2.6), the final fusion kernel elegantly combines facial and speech components:

$$ICKA_{Kernel} = d_{face} k_{face}(x_i, x_j) + d_{speech} k_{speech}(x_i, x_j). \quad (2.10)$$

The $ICKA_{kernel}$ is used as the input kernel of the SVM. This kernel formulation serves as the foundation for SVM classification, with hyperparameter optimization yielding optimal values of $C = 10$ and $\gamma = 1$ through extensive cross-validation. The resulting fusion framework effectively captures both linear and non-linear relationships between modalities while maintaining computational efficiency.

2.5.3.3 Data Fusion Performance Analysis

The performance evaluation of the data fusion approaches revealed significant insights into their comparative efficacy for QoE estimation. The experimental results, presented in Table 2.15, demonstrate the performance metrics achieved by the SVM classifier when implementing both PCA and ICKA data fusion techniques. The PCA-based solution achieved a mean accuracy of 0.84, which, while substantial, fell short of the ICKA-based method by approximately 9%. A detailed analysis of the recall metrics revealed that both approaches achieved comparable performance for the extreme rating classes (1 and 5). However, the PCA-based method exhibited less stability and diminished performance across the intermediate classes. In contrast, the ICKA-based approach maintained consistently high recall values exceeding 0.90 across all classes. The precision and F1-score metrics further reinforced the superior performance of the ICKA method, which demonstrated reduced misclassification issues while maintaining metrics consistently above 0.80. This robust performance validates the ICKA-based SVM's capability to reliably predict across the full spectrum of the five-point quality rating scale. When comparing these results with the individual facial- and speech-based SVM models (documented in Tables 2.9 and 2.13), the data fusion approach demonstrated marked improvements in estimation performance. The individual models exhibited notable weaknesses in predicting the middle class (3), showing pronounced bias toward the external classes. Specifically, the best-performing NRS SVM method achieved substantially lower recall scores compared to the ICKA-based approach, with particularly poor performance for classes 2, 3, and 4 (recall scores below 22%, 16%, and 22% respectively). To provide additional validation of the estimation accuracy, we conducted correlation analysis between the actual ACR scores and those estimated by both fusion approaches. The PCC, which quantifies linear correlation on a scale from -1 to 1, revealed strong positive correlations for both methods. The ICKA-based solution achieved a PCC of 0.984, surpassing the PCA-based approach's value of 0.923. The RMSE, which measures prediction accuracy through quadratic mean differences between true and estimated values, further demonstrated the superior performance of the ICKA method. The ICKA-based solution achieved an RMSE of 0.255, approximately half that of the PCA-based method's 0.561. These comprehensive performance metrics establish the ICKA-based solution as the most effective approach for estimating individual ACR scores using combined speech and facial features. While both fusion methods demonstrated strong correlation with actual scores, the ICKA approach's superior accuracy and consistency across all evaluation metrics validate

Table 2.15: QoE estimation performance achieved by the SVM model trained with facial and speech features fused with PCA and ICKA techniques.

Data fusion technique	Performance metric	ACR score				
		1	2	3	4	5
PCA	Mean Acc.			0.84		
	Precision	0.95	0.74	0.60	0.93	0.95
	Recall	0.90	0.85	0.71	0.86	0.90
	F1-score	0.93	0.78	0.68	0.87	0.93
ICKA	Mean Acc.			0.93		
	Precision	0.85	0.83	0.82	0.95	0.87
	Recall	0.92	0.92	0.91	0.99	0.93
	F1-score	0.87	0.85	0.80	0.95	0.88

its selection as the optimal solution for QoE estimation in this context.

2.5.4 Comparison with the state-of-the-art

A comprehensive comparison of the proposed methodology's performance against existing state-of-the-art approaches is presented in Table 2.16. The evaluation metric chosen for this comparative analysis is mean accuracy, providing a standardized basis for performance assessment across different methodological approaches. Our proposed methodology represents a significant advancement in QoE estimation by uniquely integrating both facial and speech features, achieving a mean accuracy of 0.93. This comprehensive approach distinguishes our work from existing methods in the literature, which typically focus on single modality analysis. When evaluating our method using only speech features, as detailed in Section 2.5.2, it achieved an accuracy of 0.86, demonstrating the robust performance of even the individual components of our approach. In comparison, the MLQoE methodology proposed by [100] relies exclusively on network conditions as input parameters (as specified in Table 2.2), utilizing multiple ML algorithms for QoE estimation. This approach achieved an accuracy of 0.76, with its relatively lower performance potentially attributable to the absence of speech-related feature integration in the estimation process. The SpeechQoE model introduced by [26] implements a different approach by transforming speech-based cues into QoE scores. This model processes time-frequency domain speech spectrograms, extracting implicit features through multiple levels of non-linear operations. In our comparative analysis, we utilized the time-frequency domain speech spectrograms from the OS files described in Section 2.5.2. The SpeechQoE model achieved an accuracy of 0.81. However, it is important to note that this model was developed and trained under less challenging network conditions compared to our study. Specifically, the PLR in their experiments ranged from 0 to 3%, and delay values were limited to 0-500 ms, with no consideration of

Table 2.16: Comparison with the state-of-the-art in terms of the mean accuracy.

Method	MLQoE [100]	SpeechQoE [26]	Our (speech)	Our (facial & speech)
Accuracy	0.76	0.81	0.86	0.93

jitter effects. These less demanding conditions may explain the performance differential observed when compared to our methodology. This comparative analysis demonstrates the superior performance of our integrated approach, highlighting the advantages of combining multiple modalities in QoE estimation. The results underscore the effectiveness of our methodology in handling more challenging network conditions while maintaining high estimation accuracy.

2.6 Key Takeaways and Future Directions

In this chapter 2, this doctoral research thesis has advanced the field of quality assessment in real-time communications through the innovative integration of facial expressions and voice features to predict user QoE during WebRTC video calls. The research progressed through four major phases, each building upon previous findings to create a comprehensive framework for QoE assessment.

QoE-1 (see in the Section 2.2) [10] established the foundational understanding through rigorous subjective assessment of network impairments' impact on user experience. This phase investigated how delay, jitter, and packet loss affect conversation quality and user emotional responses in WebRTC applications. Through controlled experiments with 20 participants engaging in structured video conversations, we quantified the relationships between network conditions and perceived quality, while also examining which media component (audio or video) demonstrates greater sensitivity to network impairments.

QoE-2 (see in the Section 2.3) [11] extended the analysis to application-layer data, focusing on WebRTC session parameters collected through the `webrtc-internals` tool. This investigation identified key performance indicators at the application level that correlate with user experience, employing statistical validation through ANOVA and PCC analysis. The findings enabled the development of machine learning models capable of estimating QoE from readily available application metrics.

QoE-3 (see in the Section 2.4) [12] focused on facial expression analysis, employing both traditional machine learning approaches and advanced FCN architectures. This phase extracted and analyzed facial AUs and developed novel preprocessing methods to reduce data dimensionality while preserving quality-relevant information. The comparative analysis of different machine learning approaches revealed the superiority of specific feature selection and model architectures for facial-based QoE estimation.

QoE-4 (see in the Section 2.5) [13] culminated in the integration of facial and

speech features through sophisticated data fusion techniques. This phase employed the ICKA method to combine facial and speech datasets effectively, achieving a remarkable accuracy of 0.93. This significant improvement over single-modality approaches (0.78 for facial and 0.86 for speech) validates the effectiveness of our multimodal methodology.

As a conclusion, a key advantage of our approach lies in its ability to construct personalized QoE models by capturing instinctive user reactions through both facial expressions and voice characteristics. The comprehensive analysis of 129 statistically significant features (111 voice-related and 18 facial-related) provides deeper insights than conventional models, which often overlook human influencing factors due to data collection challenges. Future research directions include extending this methodology to broader application scenarios, as we hypothesize that the correlation between our identified features and QoE may exhibit application-independent characteristics. By incorporating application-specific parameters into the system, we aim to develop a generalized model achieving comparable performance across diverse use cases. To address varying parameter sets across different application configurations, we plan to investigate Multi-view learning approaches, which offer improved generalization efficiency through multiple viewpoint analysis. This research has established a foundation for multimodal QoE assessment in real-time communications, opening new avenues for personalized service delivery and quality management in next-generation communication systems.

Furthermore, our methodological framework exhibits alignment with standardization efforts currently being pursued by the Moving Picture, Audio and Data Coding by Artificial Intelligence (MPAI) consortium. Specifically, our multimodal feature extraction and fusion techniques share conceptual foundations with MPAI's ongoing work in context-aware audiovisual analysis frameworks. The feature selection methodology and statistical validation procedures employed in our research could potentially inform MPAI's emerging standards for emotion recognition and quality assessment in conversational contexts. This connection presents opportunities for bidirectional knowledge transfer, whereby our empirically validated approach to multimodal QoE estimation could contribute to standardization efforts, while simultaneously benefiting from the evolving frameworks being developed within the consortium.

In addition, it could be integrated into dynamic control mechanisms of 5G/Sixth Generation Technology (6G) networks through strategic feature distribution across network layers. At the terminal level, lightweight feature extraction could capture fundamental facial and speech characteristics, while edge nodes could perform intermediate processing, and cloud environments could execute sophisticated fusion models. This distributed architecture would enable real-time QoE-based network adaptation while optimizing computational resource utilization. Although our approach was developed for WebRTC, it maintains transferability to general streaming protocols like HLS with appropriate modifications. These would include the extraction of client-side features, periodic transmission through side channels, and adaptation of

fusion models to incorporate streaming-specific parameters. Such implementations would require recalibration of feature importance weights and consideration of the asynchronous nature of streaming compared to real-time communication.

Chapter 3

Balancing Quality and Sustainability in Video Streaming

The delivery of Internet content has expanded considerably in recent years. Video traffic has grown substantially, accounting for 82% of all consumer Internet traffic in 2022, up from 75% in 2017 [101]. This dramatic increase stems from multiple factors: the proliferation of video streaming services like live streaming, video-conferencing and video-on-demand; the availability of higher-resolution content including 4K, 8K, and Augmented Reality (AR)/Virtual Reality (VR); and the ability to consume video across diverse end-devices such as TVs, PCs, and mobile devices. As video consumption has grown, so too have end-user expectations around quality. This has driven the development and implementation of QoE-aware video streaming management approaches [37]. QoE refers to the degree of delight or annoyance experienced by users of an application or service [23]. QoE-aware management aims to prevent and handle potential interruptions in video delivery caused by network fluctuations and congestion, while delivering the highest possible quality content with minimal delay. However, these approaches rarely consider the energy consumption implications across the video streaming delivery chain. The negative environmental impact of ICT through increased electricity usage and resulting carbon dioxide (CO_2) emissions has sparked research interest in developing more sustainable "Green Internet" solutions. An ideal scenario would balance device energy constraints and resource limits while still meeting the QoS requirements for enjoyable video viewing and enhanced QoE. Despite rapid advancement in video streaming services, energy consumption remains a critical concern requiring improved efficiency. Given limited research examining this challenge, we investigated potential trade-offs between sustainable video streaming and QoE.

This chapter first reviews existing literature on energy consumption in online video streaming services (Section 3.1). Then, we explore key challenges and approaches for achieving sustainability while maintaining acceptable quality (Section 3.2). Our analysis suggests that limiting maximum QoE to acceptable levels, rather than pur-

suing maximum quality, can meaningfully reduce energy consumption while still satisfying users with a contribution of QoE-5 [15]. To validate these findings, we conducted a survey of 100 participants examining their willingness to accept reduced video quality when aware of potential energy and cost savings (Section 3.3). The results demonstrate that under specific conditions, particularly when quality reductions are not highly noticeable, most users would accept lower video quality to reduce energy consumption and costs.

3.1 Energy Consumption of Video Streaming

The electricity consumption of video streaming can be attributed to four main elements [102]: data centers, data transmission networks, end devices, and consumer behavior. Table 3.1 summarizes the current possible actions for sustainable video streaming that can be taken by providers and consumers based on the literature.

3.1.1 Data Centers and Data Transmission Networks

Regarding energy consumption of data centers and data transmission networks, several studies have analyzed the CO_2 emissions across the computing continuum of cloud, edge, and fog resources. Research has shown that modern edge devices like Raspberry Pi 4B and Jetson Nano emit up to six times less carbon during neural network training while achieving comparable performance to cloud and fog computing applications [103]. This suggests that offloading applications with soft execution time constraints, such as video-on-demand, data filtering, and small dataset model training, to edge devices can reduce energy costs and carbon footprint while maintaining adequate performance and QoE. A cloud-based application proposed in [104] analyzes video streams from camera networks connected through 5G, demonstrating that edge computation could reduce CO_2 emissions by up to 50%. Similarly, mobile edge caching and service offloading from macro to smaller cells have proven highly energy-efficient for future mobile network topologies [105], potentially reducing energy consumption by up to 81% for video streaming and 83% for virtual reality applications. These findings confirm that edge computing helps lower network loads and energy usage while preserving or improving end-to-end QoS and QoE by reducing application-to-consumer latency. Technical regulations for reducing energy consumption include restricting autoplay systems, implementing energy-efficient systems, and transitioning to renewable energy sources [102]. Base stations can be equipped with energy harvesting components to collect energy from wind and solar sources for storage and data transmission [106]. A decentralized resource allocation algorithm for (Green Mobile Edge Computing (GMEC)) reduces energy waste without compromising performance [107]. Time-varying renewable energy sources integrated into GMEC improve mobile client QoE while optimizing backhaul traffic in dynamic adaptive video streaming over HTTP [108]. Economic regulations

like transit fees on data traffic volumes can incentivize content providers to optimize QoE through data compression and management approaches. Studies have explored trade-offs between encoding complexity and efficiency [109], showing that lower-power encoding schemes can be used when quality improvements are negligible. Cloud-based adaptive bitrate encoding with renewable energy sources has demonstrated 14-28% energy cost reduction with minimal QoE impact [110]. Network softwarization through Software-defined networking (SDN) can also achieve energy savings while maintaining performance by reducing fog traffic latency and data center carbon footprint [111].

Table 3.1: Current possible actions to be taken from the provider’s and consumer’s side for sustainable video streaming.

Energy Consumption Element	Provider/Producer/ Government Action	Consumer Action
Data center	• Offloading contents/applications to the edge • Switching to renewable energy sources • Energy-efficient algorithms for GMEC	• Prefer providers attentive to sustainability
Data transmission network	• Setting transit fee on quantity data traffic • Data compression/ encoding algorithms • Network softwarization (e.g., SDN)	• Prefer energy-efficient network • No autoplay
End-device	• Produce energy-efficient devices	• Prefer small size (mobile) devices • Prefer low-resolution videos
Consumer behavior	• Transit fee on quantity data traffic • Reduce the number of consumers	• Reduce the time watching online video contents • Prefer watching videos with other people than alone

3.1.2 End-device

Multiple studies have identified the end device as the largest contributor to energy consumption in the video streaming chain [102, 112–114]. For instance, one hour

of Netflix streaming consumes approximately 76.9 Wh on average, with 70% consumed by the device itself, 25% by transmission, and only 5% by the data center [113]. These findings emphasize the critical importance of understanding and optimizing energy consumption at the end-device level. The device size plays a crucial role in energy consumption levels. Research has shown that smaller devices, both in terms of physical size and screen resolution, consume significantly less energy. A striking example is that a 4K TV can consume more than ten times the energy of a smartphone [113]. This principle extends to the environmental impact across the device lifecycle, with smaller devices generally having a lower environmental footprint [114]. A comprehensive survey conducted by [114] revealed distinct patterns in device-specific (Global Warming Potential (GWP)). Laptops and PCs showed the highest GWP when used for viewing content on free platforms like YouTube. Smart TVs demonstrated the greatest GWP for subscription-based services such as Netflix, while smartphones had the highest GWP impact for social media video consumption on platforms like TikTok. While smartphones and mobile devices generally consume the least energy individually, their collective impact on CO₂ emissions remains substantial due to the vast number of mobile subscribers and extensive streaming hours. This reality has prompted research into QoE-aware energy-saving approaches for mobile multimedia streaming [115, 116]. Studies have shown that modifying default adaptive bitrate policies can achieve energy savings of 10-30%, with only minimal impact on video quality [115]. Network choice also influences energy consumption patterns, with Wi-Fi generally enabling lower energy consumption compared to High Speed Packet Access (HSPA) and Long Term Evolution (LTE), which show similar consumption levels [116]. These findings point to two parallel paths forward: device manufacturers must focus on developing more energy-efficient devices that maintain or improve performance while reducing energy consumption, while consumers should be encouraged to favor smaller devices equipped with energy-efficient network interfaces when appropriate for their viewing needs.

3.1.3 Consumer behavior

The consumer behavior has a significant role in video streaming energy consumption, though it remains challenging to quantify. To collect quantitative data on consumer behaviors like streaming duration, device selection, and resolution choices, Suski et al. [114] conducted a comprehensive survey. Their results revealed that online video streaming has substantial environmental relevance and is heavily influenced by specific user behaviors. The study found that over 50% of average energy consumption from streaming can be attributed to paid platform content consumption, primarily due to longer streaming durations, along with users' preferences for larger devices such as laptops, PCs, and smart TVs. The higher energy consumption associated with paid platforms was further validated by [117], which analyzed CO₂ emissions across four major social media platforms: Netflix, YouTube, Facebook, and TikTok. The analysis considered platform activity, video data size, and user base metrics.

Table 3.2: End-device data.

Device	Standard resolution [p]	Bitrate br [kbps]	Data traffic R_i [GB/h]	Load P_i [W]
Smartphone	360p	700	0.315	6
Tablet	480p	1250	0.5625	7
Laptop/PC	720p	4125	1.856	32
Smart TV	1080p	6750	2.168	200

Netflix emerged as the highest CO₂ generator at 983 grams per hour, attributed to its high-resolution content, substantial user base, and extended viewing sessions. This was followed by TikTok at 327 grams per hour, YouTube at 164 grams per hour, and Facebook at 16 grams per hour. These findings emphasize how platform user numbers and viewing duration significantly impact energy consumption. Madlener et al. [102] examined potential power savings in European video streaming from 2020 to 2030. Their analysis incorporated seven key descriptors: subscriber numbers, CO₂ emission intensity, electricity intensity across data centers, networks, and end-user devices, streaming duration, and hourly data flow based on resolution choices. The researchers modeled electricity consumption using three scenario projections: Business-as-usual (BAU) representing current trend continuation, Gray depicting a pessimistic outlook with constant CO₂ emissions exceeding BAU levels, and Green presenting an optimistic future with enhanced energy efficiency, reduced emissions, and more environmentally conscious consumer behavior. Their model demonstrated that European regulatory interventions outlined in the Green scenario could substantially reduce electricity consumption and CO₂ emissions in the video streaming chain. However, achieving these environmental benefits requires consumers to adopt more energy-efficient behaviors, including reducing viewing hours, selecting smaller devices, and opting for lower video resolutions. These findings highlight the crucial role of consumer choices and behavior patterns in determining the environmental impact of video streaming services.

3.2 Trade-off between Sustainability and QoE

This research investigates the possibility of finding a trade-off between sustainability and QoE. Through analysis, limiting the maximum QoE to an acceptable QoE emerges as a potential solution to reduce energy consumption while maintaining consumer satisfaction. A survey conducted to examine people's willingness to reduce video streaming quality when aware of potential energy savings partially validates these findings. The results demonstrate that most consumers accept lower video quality within certain bounds if it enables reduced energy consumption and cost savings. While the previous section illustrated how different components of the video streaming chain contribute to energy consumption and CO₂ emissions, achiev-

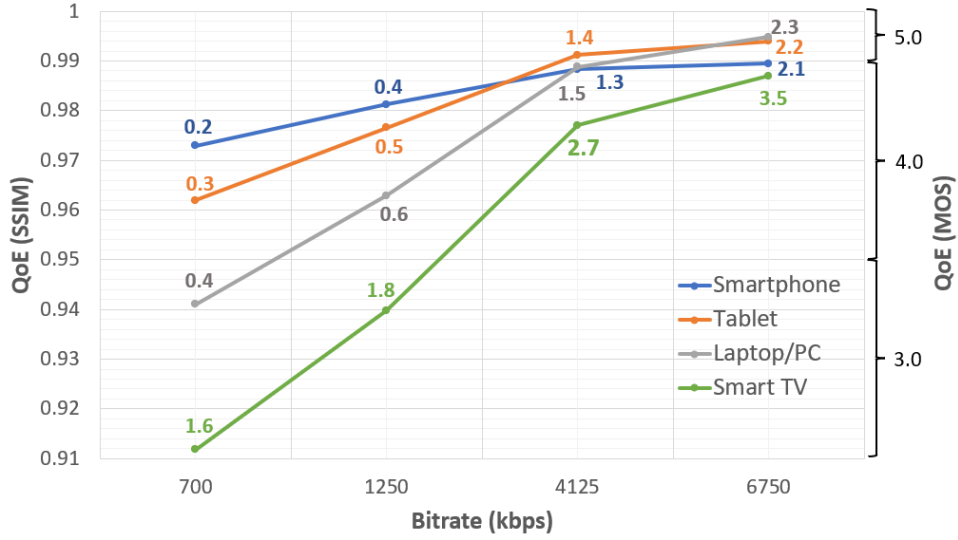


Figure 3.1: QoE (SSIM and MOS) vs. bitrate. The respective values of electricity consumption Q (kWh) are also provided on the graph with colored numbers.

ing reduced energy demand and emissions while minimizing impact on end-user QoE remains a significant challenge. For data centers and transmission networks, some energy-saving solutions have already been proposed and deployed that aim to preserve service quality. However, as established earlier, end devices and consumer behavior represent the largest energy costs in video streaming, making these elements particularly challenging to address. The most relevant findings from studies discussed in Section 3.1 indicate that meaningful energy savings would require reducing subscriber numbers, decreasing streaming hours, favoring smaller devices, and lowering video resolution. Regarding subscriber and streaming hour reductions, these could occur naturally if consumers become aware of video streaming’s environmental impact. Consumers may choose to watch less or cancel subscriptions, though likely only a minority would take such extreme measures. Additionally, such reductions would impact content provider revenue and thus face resistance. Nevertheless, this research examined consumer willingness to reduce streaming hours through a survey presented in Section 3.3. The use of smaller devices and lower resolutions may negatively impact user QoE since video quality significantly influences this metric. Therefore, research must explore the balance between perceived QoE and device energy consumption. Hossfeld et al. [118] examined potential trade-offs between QoE and CO_2 emissions for video-conferencing services using a logarithmic QoE model mapping video bitrate to MOS. Their observations showed MOS scores of 4.75 and 4.5 required bitrates of 11 Mbps and 4.5 Mbps respectively. Since higher bitrates increase energy consumption, significant energy savings become possible by providing good quality (MOS around 4) rather than maximum QoE. This aligns with Fiedler et al. [119], who defined “QoE-aware sustainable throughput” as the

maximum throughput maintaining acceptable QoE disturbances from delivery problems. While they considered throughput rather than bitrate, both approaches aim to save energy while ensuring good rather than maximum QoE. Similarly, Herglotz et al. [120] measured PC and laptop power consumption during video streaming, finding potential energy efficiency improvements of at least 3.5% while maintaining high QoE through optimized video parameters. These studies demonstrate that increased service performance does not necessarily yield proportionally better perceived QoE. For sustainable video streaming, considering "acceptable QoE" - the quality level acceptable to environmentally-conscious consumers - may prove more practical than pursuing maximum QoE. However, acceptability varies based on:

1. context (e.g., which platform the video is watched), device characteristics (small-size screens can provide good QoE even with medium-quality video content)
2. content type (e.g., preferred video genre)
3. consumer expectations (e.g., a consumer used to watching high-quality video contents would be more quality pretender)
4. service cost (e.g., the higher the cost, the higher the consumer's expectations)

The streaming platform, screen size, video genre preferences, quality expectations, and price all influence what users consider acceptable quality. This complex relationship between acceptable QoE and sustainable streaming requires investigation across multiple dimensions. The following section examines how different end devices impact perceived QoE and electricity consumption.

3.2.1 Video bitrate vs. Electricity consumption vs. QoE

This section examines the interrelationship between video bitrate, electricity consumption, and perceived QoE for video streaming services accessed through four distinct end-devices: smartphone, tablet, laptop/PC, and smart TV. To estimate the electricity consumption for device streaming, we employ the equation proposed in [114]:

$$Q_i = t_i \cdot (P_i + R_i \cdot \rho), \quad (3.1)$$

where Q_i represents the electricity consumption (kWh) of device i , t_i denotes the streaming hours over a week for device i , P_i indicates the power load (kW) of device i , R_i represents the data traffic (GB/h) for a given bitrate, and $\rho = 0.1$ kWh/GB signifies the electricity intensity of data traffic, encompassing both data transmission network and data center consumption [114]. Table 3.2 presents the comprehensive end-device data. The power load data is derived from [114], while the bitrate for each resolution corresponds to YouTube's median suggested bitrate

for specific resolutions¹. For our simulations, we established $t_i = 7$ uniformly across all devices i to facilitate direct comparison of electricity consumption patterns across different devices over identical time periods. To estimate the perceived QoE based on video bitrate, we utilize the objective QoE model introduced in [121]:

$$QoE = a \cdot br^b + c, \quad (3.2)$$

where br represents the bitrate, and a , b , and c are resolution-specific regression coefficients [121]. The QoE is quantified using the SSIM (Structural SIMilarity) metric, where a MOS of 5 corresponds to $SSIM \geq 0.99$, a MOS of 4 to $0.95 \leq SSIM < 0.99$, and a MOS of 3 to $0.88 \leq SSIM < 0.95$ [122]. Through the application of Eq. (5.2) and Eq. (3.2), we compute both energy consumption and perceived QoE across the four end-devices as functions of bitrate. Fig. 3.1 illustrates the relationship between bitrate and QoE in terms of SSIM and MOS, with corresponding electricity consumption Q (kWh) values indicated by colored numbers. As anticipated, the Smart TV demonstrates the highest electricity consumption among all devices, while other end-devices exhibit lower and relatively comparable consumption patterns. The data reveals a direct correlation between video bitrate and device electricity consumption, primarily due to increased data traffic requirements for higher-resolution video encoding. The presented electricity consumption values reflect a seven-hour weekly streaming period, with both consumption levels and inter-device consumption gaps expanding proportionally with increased streaming hours, underscoring the significant impact of consumer behavior on sustainability. Regarding QoE, a crucial question emerges: to what extent does increasing allocated resources (in terms of resolution and encoding bitrate) correspond to enhanced user-perceived QoE? For smartphones and tablets, the two lowest bitrates achieve a MOS around 4, while the two highest bitrates reach a MOS around 5, suggesting that bitrate increases beyond certain thresholds result in wasteful network bandwidth consumption and increased energy usage without meaningful QoE improvements. Similar patterns emerge for laptop/PC usage, with the notable exception of the lowest bitrate achieving a MOS below 4. The dynamics shift significantly with larger end-devices, particularly Smart TVs, where their substantially wider screens demonstrate different quality-bitrate relationships. The lowest two bitrates achieve only sufficient quality (MOS of 3), while good to excellent QoE is attained at the highest two bitrates. However, the perceptible difference between the two highest bitrates may vary among consumers based on their individual acceptability profiles. These findings align with previous research [118–120], demonstrating that improvements in system performance do not necessarily yield proportional increases in perceived QoE. This complex relationship between electricity consumption, system performance, and perceived QoE warrants further investigation to optimize resource allocation while maintaining user satisfaction.

¹<https://support.google.com/youtube/answer/2853702>

3.3 Are users willing to save energy?

To understand consumers' attitudes toward energy conservation in online video streaming, we conducted a crowdsourcing survey using the Amazon Mechanical Turk (MTurk) platform². The survey targeted workers meeting specific criteria: a 98% approval rating to ensure reliable responses, English as their native language to maximize reach, and PC/laptop as their primary device since some questions required viewing videos on larger screens. Each participant spent approximately 10 minutes completing the survey and received compensation of 0.50\$. The survey gathered responses from 100 participants (57 males, 43 females) with a mean age of 36.5 years (ranging from 17 to 72 years). The geographical distribution included 28 respondents from Asia, 37 from Europe, and 35 from North/Central America. Initial questions focused on video streaming habits, revealing that most participants preferred larger devices for streaming: 48 used laptop/PC and 28 used TV, compared to only 18 using smartphones and 6 using tablets. The sample demonstrated high familiarity with video streaming services, with 88 participants streaming daily - 51 for 0-2 hours and 37 for 2-4 hours. Notably, 72 participants relied on automatic resolution settings, while 28 manually adjusted settings based on screen size. This tendency toward automatic settings suggests potential energy waste, as Over-The-Top providers like Netflix typically deliver the highest possible quality to reduce user churn, regardless of necessity. Participants viewed a 10-second clip from Big Buck Bunny³ encoded at three different resolutions (720p, 480p, 360p). They rated the perceived QoE using the 5-level (1-Bad, 2-Poor, 3-Fair, 4-Good, and 5-Excellent) single discrete ACR scale, following ITU-R Rec. P.800 [82]. The MOS results showed a direct correlation between resolution and user satisfaction: 720p achieved 4.37, 480p reached 3.85, and 360p scored 3.52. The survey included three key environmental concern questions:

1. (QE1) Participants' willingness to reduce video streaming time for environmental reasons.
2. (QE2) Their openness to reducing video quality for energy and cost savings.
3. (QE3) Specific willingness to accept quality reductions with known energy savings percentages.

Responses to QE1 and QE2 showed similar patterns, with approximately 40% agreeing, 20% disagreeing, and 15% strongly disagreeing. Agreement and neutral responses were minimal, around 10% each. These results indicate that at least half the participants showed environmental consciousness and willingness to modify their streaming habits for sustainability.

QE3 explored three specific scenarios:

²<https://www.mturk.com/>

³<https://peach.blender.org/>

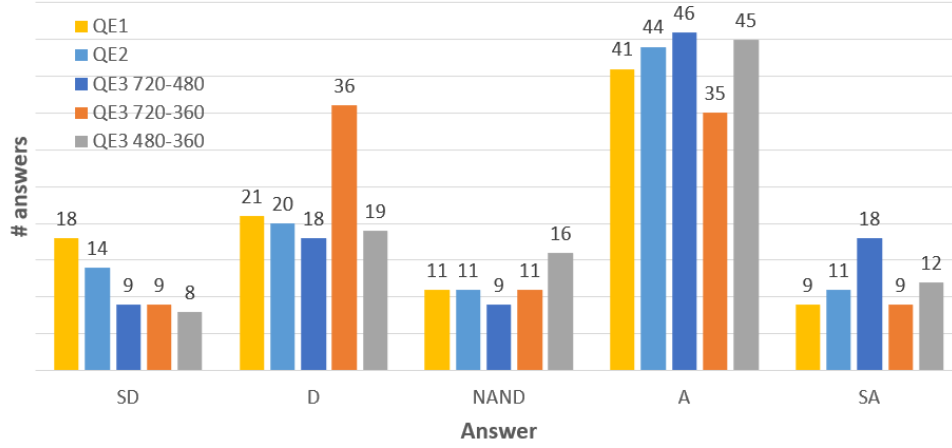


Figure 3.2: Answers for environmental concern. SD: Strongly Disagree, D: Disagree, NAND: Neither Agree Nor Disagree, A: Agree, SA: Strongly Agree.

1. Reducing quality from 720p to 480p (30% energy savings)
2. Reducing quality from 720p to 360p (75% energy savings)
3. Reducing quality from 480p to 360p (45% energy savings)

All videos were displayed in fixed-size boxes with disabled full-screen mode. Energy savings calculations utilized equations from Section 3.2. Results showed nearly 50% agreement for scenarios 1 and 3, while scenario 2 divided respondents evenly between agreement and disagreement (35% each). This variation likely stems from perceived quality differences: the shift from 720p (MOS 4.37) to 480p (MOS 3.85) maintained acceptable quality while offering energy savings. However, the more dramatic quality reduction to 360p (MOS 3.52) met greater resistance despite higher energy savings. These findings demonstrate that many users are willing to balance sustainability and QoE, each determining their personal acceptable quality threshold when aware of potential energy and cost savings. The results suggest promising opportunities for implementing energy-efficient streaming options that maintain user satisfaction while reducing environmental impact.

3.4 Outcomes and Discussion

The challenge of sustainable video streaming lies in finding an effective balance between limiting energy consumption and CO_2 emissions while maintaining satisfactory user QoE. Through a comprehensive analysis, this research has identified that end-device usage and consumer behavior represent the most significant contributors to energy costs in the video streaming delivery chain. A clear correlation exists between device size and energy consumption: larger devices with higher screen

resolutions inherently consume more power. Additionally, these wider screens typically demand higher resolution video content, resulting in increased data traffic and consequently higher energy usage across the entire delivery infrastructure. This study's analysis has revealed that focusing on acceptable rather than maximum QoE presents a viable solution for reducing energy consumption while maintaining user satisfaction. This finding is supported by the observation that increases in system performance do not necessarily translate to proportional improvements in perceived QoE. A key validation of this concept came through survey results that examined users' willingness to accept reduced video streaming quality when aware of potential energy savings. The findings demonstrated that under specific conditions, particularly when quality reductions remain relatively imperceptible, most users would willingly accept lower video quality in exchange for reduced energy consumption and cost savings. These insights suggest promising opportunities for the development of new energy management strategies that prioritize acceptable QoE rather than maximum performance. Such approaches could help address the growing environmental impact of video streaming services while ensuring continued user satisfaction. The findings open new avenues for research into energy-efficient video delivery systems that balance performance requirements with sustainability goals. By understanding and leveraging the relationship between perceived quality and energy consumption, service providers can work toward more sustainable solutions that align with both user expectations and environmental responsibilities. This research contributes to the broader discussion of sustainable digital services and provides a foundation for future work on energy-sensitive video streaming optimization.

To further explore these initial findings and deepen our understanding of the relationship between sustainable video delivery and user experience, we conducted a comprehensive subjective evaluation study detailed in Chapter 4 with a contribution of QoE-6 [16]. This follow-up investigation specifically examined how video resolution, various types of luminance configurations, and different end devices impact both QoE and energy consumption in video streaming services. Through this real-world experimental setup, we sought to validate and expand upon our theoretical framework while gathering concrete empirical evidence about the practical implications of different video delivery configurations on both user satisfaction and energy efficiency.

Chapter 4

Are Quality and Sustainability Reconcilable? A Case Study on Video Quality, Luminance and Resolution

The pervasive impact of ICT services on energy consumption and $CO_2!$ ($CO_2!$) emissions has catalyzed extensive research initiatives focused on developing the Green Internet paradigm, which endeavors to design environmentally sustainable ICT services. This transformation necessitates a thorough understanding of the relationship between QoE and sustainable video streaming services.

The present research with a contribution of the thesis QoE-6 conducts a comprehensive subjective assessment to examine the intricate interplay between video resolution, luminance configurations, and device characteristics in relation to both QoE and energy consumption patterns in video streaming services [16]. The investigation employs rigorous statistical methodologies, specifically Analysis of Variance and t-test models, to analyze subjective data and identify the primary factors influencing QoE and energy consumption. The findings indicate that under specific operational conditions - including ambient illumination levels, device backlight luminance settings, and display screen dimensions - users can be effectively guided towards an optimal compromise between acceptable QoE and sustainable choices. This is particularly evident in scenarios where increased energy expenditure fails to yield proportional enhancements in user experience.

Contemporary research initiatives predominantly focus on algorithmic optimization for renewable energy utilization, computational complexity reduction, data storage efficiency, and device power management [105, 107]. However, these approaches frequently overlook the user impact dimension, which remains intrinsically linked to service delivery outcomes.

The implemented modifications must carefully consider the extent of user experience alterations in relation to potential environmental benefits. This research specifically addresses video streaming services, a critical focus area given the substantial increase in video traffic, which experienced 24% growth in 2022 and currently constitutes 65% of total Internet traffic [2]. To establish a comprehensive understanding of how various video streaming parameters affect energy consumption and user QoE, we designed and executed a detailed subjective assessment protocol. The investigation encompasses multiple key parameters, with particular emphasis on video resolution, which significantly impacts both user QoE and energy consumption. The relationship between video resolution and energy consumption is particularly noteworthy, as the end-to-end video streaming chain’s electricity consumption correlates directly with data traffic volume, which increases proportionally with video resolution and bitrate [114]. The research framework incorporates three distinct end-devices (TV, laptop, and smartphone) and examines three luminance categories: device backlight luminance, ambient luminance, and content luminance. This comprehensive approach acknowledges that user visual perception is substantially influenced by contextual factors, including display dimensions, device backlight settings, and ambient lighting conditions during video consumption [123]. Moreover, display size and backlight luminance represent significant contributors to overall energy consumption, particularly in larger display configurations.

The experimental methodology comprised 16 distinct test conditions, systematically varying the aforementioned parameters, with evaluations conducted by 25 participants. The analysis proceeded in two phases: first, examining the subjective data to identify primary QoE influencing factors, and second, comparing QoE outcomes between test conditions selected according to specific criteria. These criteria included: comparing conditions with differing energy consumption profiles based on user-controllable parameters (backlight luminance and video resolution), while maintaining consistency in non-user-controllable parameters. The statistical analysis yields compelling evidence that under specific conditions, users can be effectively guided toward environmentally sustainable choices without significant compromise to their viewing experience, as increased energy consumption does not necessarily correlate with proportional QoE improvements.

This paper proceeds with the following structure: Section 4.1 delineates the experimental scenario, while Section 4.2 details the subjective assessment methodology. Section 4.3 provides a thorough analysis of the experimental outcomes, and Section 4.4 presents the concluding observations and future research directions.

4.1 Scenario Settings

The investigation of luminance configurations and video resolutions’ impact on QoE and electricity consumption in video streaming across various device types represents a critical research focus. This significance stems from both resolution and lumi-

nance settings being substantial contributors to global CO₂ emissions, particularly when considering the cumulative hours of video streaming consumption. To conduct a comprehensive analysis, the research methodology incorporated two distinct video resolutions, each associated with corresponding bitrates, across three device categories differentiated by screen dimensions: Television (TV), Laptop (LP), and Smartphone (SP).

The experimental design aimed to present video content at both the native resolution of each device and at a reduced yet acceptable resolution to evaluate perceptual quality distinctions. A noteworthy consideration in this context is the prevalent tendency among both participants and video players to overestimate necessary video quality requirements, particularly for mobile devices. This overestimation phenomenon leads to the deployment of unnecessarily high bitrates, resulting in inefficient utilization of network bandwidth during streaming operations.

The research framework incorporated three distinct categories of luminance: Ambient Luminance (AL), device Backlight Luminance (BL), and Content Luminance (CL). AL, measured in lux (lx), encompasses any illumination not originating from the device screen [123], including both natural sources (e.g., solar radiation) and artificial illumination (e.g., electric lighting). BL refers to the screen-generated luminance, which can be adjusted through predefined discrete steps ranging from minimum to maximum brightness. CL represents the pixel-level luminance variations inherent in the video content, fluctuating between dark and bright scenes. While all three luminance types potentially influence user QoE, BL and AL demonstrate the most significant impact on electricity consumption. To evaluate these parameters, a subjective quality assessment was conducted wherein participants evaluated video sequences encoded at varying resolutions across the three device types under different combinations of the luminance variables. The experimental conditions were systematically defined and implemented according to the test conditions (TCs) outlined in Table 4.1. This methodological approach enabled a comprehensive examination of the interrelationships between video quality parameters, device characteristics, and user perception.

4.1.1 Devices and Content Resolution

The experimental framework incorporated three distinct device categories with varying display specifications to ensure comprehensive evaluation across different viewing contexts. The primary display device consisted of a Samsung 60" Flat Smart Serie 6 JU6800 television, featuring an Ultra High Definition (UHD) 4K screen with a resolution of 3840×2160 pixels and 73 pixels per inch (ppi). This TV implementation utilized edge LED backlight technology and presented substantial viewing dimensions of 136.36 cm in width and 78.95 cm in height. For the intermediate viewing scenario, we employed an ASUS N56V8 LP equipped with a 15.6" Liquid Crystal Display (LCD) display supporting Full High Definition (1920×1080) resolution. The computational capabilities of this device were ensured through an Intel i7-3630QM

Table 4.1: The 16 Test conditions (TCs). BL and RES values depend on the end device.

TC	BL (lx)			AL (lx)	RES			CL
	TV	LP	SP		TV	LP	SP	
TC1	300	400	200	0	4K	FHD	FHD	H
TC2					FHD	HD	HD	
TC3					4K	FHD	FHD	L
TC4					FHD	HD	HD	
TC5				500	4K	FHD	FHD	H
TC6					FHD	HD	HD	
TC7					4K	FHD	FHD	L
TC8					FHD	HD	HD	
TC9	5500	4000	5000	0	4K	FHD	FHD	H
TC10					FHD	HD	HD	
TC11					4K	FHD	FHD	L
TC12					FHD	HD	HD	
TC13				500	4K	FHD	FHD	H
TC14					FHD	HD	HD	
TC15					4K	FHD	FHD	L
TC16					FHD	HD	HD	

processor operating at 2.40 GHz, complemented by 8GB of Random Access Memory (RAM) and dedicated graphics processing via an NVIDIA GT740M Graphics Processing Unit (GPU) with 2GB of video memory. The mobile viewing context was represented by a Huawei P9 Plus SP, incorporating a 5.5" LCD display with Full High Definition (FHD) resolution (1920×1080). This device's processing architecture consisted of a HiSilicon Kirin 955 chipset, supported by 4GB of RAM and graphics rendering through an ARM Mali-T880 MP4 GPU. For content evaluation, we selected the Tears of Steel (ToS) ¹ video sequence as our reference material. The spatial resolution parameters were carefully calibrated to align with each device's native display capabilities. For the TV, we implemented both 4K (3840×1714) and FHD (1920×800) resolutions, while the LP and SP configurations utilized FHD and High Definition (HD) (1280×534) resolutions. This methodological approach ensured that the highest resolution setting corresponded to each device's native display capacity, with the lower resolution representing the next standardized format below the native specification. To maintain temporal consistency across all experimental conditions, we preserved the original frame rate of 24 frames per second from the ToS source material across all device configurations and resolution settings. This standardization of temporal parameters facilitated direct comparative analysis while eliminating potential variables related to frame rate differences.

¹<https://mango.blender.org/>

4.1.2 Luminance

The experimental design incorporated three distinct types of luminance measurements: AL, BL, and CL. For ambient luminance assessment, two controlled conditions were established: a dim light environment approximating 0 lx to simulate dark conditions, and a moderate indoor lighting level of 500 lx to represent typical ambient illumination. The selection of 500 lx was methodologically justified as an intermediate value between residential indoor lighting (100 lx) and bright indoor lighting (1000 lx), following precedents established in prior research [123]. Environmental illumination control was achieved using three dimmable light sources, with precise luminance measurements obtained via the VEML7700 ambient light sensor. The measurement protocol adhered to ITU-R BT.500-14 recommendations, specifically maintaining perpendicular incident light measurement relative to device screens [84]. For device backlight luminance evaluation, the experimental protocol examined extreme brightness conditions, measuring both minimum and maximum BL settings for each display device. Objective BL measurements were conducted using the VEML7700 light sensor positioned at a standardized 5 cm distance from each screen surface. Measurement consistency was maintained by displaying an identical total-white image across all devices while eliminating confounding light sources from the testing environment. The empirical measurements revealed maximum BL values of 5500, 4000, and 5000 lx for TV, LP, and SP respectively, with corresponding minimum values of 300, 400, and 200 lx. Content luminance quantification employed the Earth Mover’s Distance (EMD) metric [124], selected for its capacity to capture both inter-pixel luminance interactions and frame-to-frame luminance discontinuities [123]. The methodological approach calculated EMD values between median video scene histograms and a reference black image histogram, with median histogram bins derived from frame-level histogram analysis. The experimental content comprised sixteen 10-second video sequences stratified into two luminance categories: eight sequences exhibiting low EMD values (range: 40-60, mean: 53.21) and eight sequences with high EMD values (range: 90-140, mean: 102.39). Each sequence underwent encoding at three distinct resolutions: 4K, FHD, and HD. The presentation order of low and high CL sequences across different test conditions was randomized to minimize order effects. A comprehensive overview of test conditions per device type is presented in Table 4.1.

4.1.3 Electricity consumption

The estimation of electricity consumption for video streaming services requires consideration of multiple components across the end-to-end delivery chain. Following the methodology proposed by [114], the overall electricity consumption can be quantified using Equation 5.2:

$$Q_i = t_i \cdot (P_i + R_i \cdot \rho), \quad (4.1)$$

Table 4.2: Estimated overall electricity consumption per week for TV, LP, and SP.

Device	RES	R_i (GB/h)	Avg BR (Mbps)	BL (lx)	P_i (W)	Q_i (kWh)
TV	4K	9.00	20	300	55.78	6.69
				5500	236.05	7.95
	FHD	2.25	5	300	53.67	1.95
				5500	243.48	3.28
LP	FHD	2.25	5	400	13.04	1.67
				4000	18.56	1.70
	HD	0.90	2	400	13.21	0.72
				4000	18.53	0.76
SP	FHD	2.25	5	200	1.85	1.59
				5000	2.46	1.59
	HD	0.90	2	200	2.98	0.65
				5000	3.19	0.65

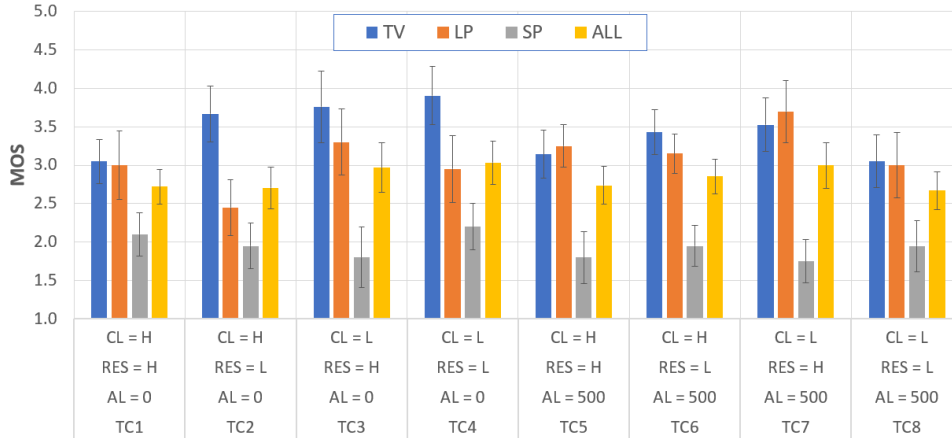


Figure 4.1: Devices' screens with BL = Min. TV: 300 lx, LP: 400 lx, SP: 200 lx. Mean Opinion Score (MOS) with 95% confidence interval (CI) for the first eight TCs (from 1 to 8).

where Q_i represents the electricity consumption in kilowatt-hours (kWh) for device i per week, t_i denotes the streaming duration in hours over a week for device i , P_i indicates the power load in kilowatts (kW) of device i , R_i signifies the data traffic in gigabytes per hour (GB/h) which varies based on video resolution and bitrate, and ρ represents the electricity intensity factor for data traffic, encompassing both data transmission network and data center consumption. Based on recent research [102], ρ is established at 0.1 kWh/GB. To accurately measure the power load (P_i) of the devices, we employed a Smart Plug TP-Link HS110 in conjunction with the TPLink Energy Monitor tool to collect power measurements. The measurements

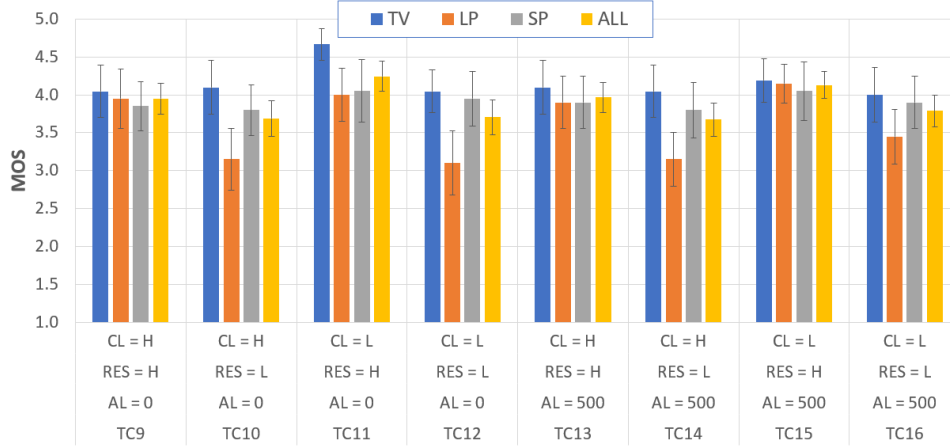


Figure 4.2: Devices' screens with BL = Max. TV: 5500 lx, LP: 4000 lx, SP: 5000. Mean Opinion Score (MOS) with 95% confidence interval (CI) for the second eight TCs (from 9 to 16).

were conducted while playing the complete Tears of Steel (ToS) video sequence, approximately 12 minutes in duration, with both minimum and maximum BL settings. For precise measurements, batteries were removed from both laptop and smartphone devices, which were instead connected directly to mains power. The experimental setup varied by device type: for the TV, 4K and FHD videos were accessed from a hard disk via USB 3.0 port using the native video application; for the LP and SP, FHD and HD videos were stored locally and played using VLC application and native video application respectively. The measurement results, summarized in Table 4.2, reveal several significant findings regarding power consumption patterns. Video resolution demonstrates minimal impact on device power load (P_i), with the exception of the SP. Interestingly, in certain scenarios, devices consume more power when processing videos encoded at resolutions below their native screen resolution, potentially due to the energy demands of upscaling algorithms. The impact of BL on power consumption shows strong correlation with screen size. The TV's power load increases more than fourfold when switching from minimum to maximum BL. The LP exhibits a more modest increase of approximately 5W (less than 50%), while the SP shows minimal variation, with the power ratio barely exceeding 1 for FHD playback and approaching unity for HD content. For overall electricity consumption calculation, we standardized $t_i = 7$ across all devices, simulating one hour of daily video streaming over a week. When considering the complete end-to-end streaming chain, video resolution emerges as the dominant factor in electricity consumption due to its direct influence on data traffic volume. The BL impact on overall consumption is primarily significant for the TV, where the larger screen size amplifies the energy requirements when transitioning from minimum to maximum brightness. The analysis reveals substantial differences in energy consumption across different viewing scenarios. Streaming 4K versus FHD content on the TV increases overall

electricity consumption by factors of 3.4 and 2.4 for minimum and maximum BL settings respectively. For both LP and SP, the transition from HD to FHD results in electricity consumption increases of 2.3 and 2.45 times respectively, independent of BL settings.

4.2 Subjective assessment

The subjective assessment methodology incorporated a diverse participant pool comprising 25 individuals, a sample size deemed sufficient according to the ITU-T Rec. G.114 recommendation. The gender distribution included 16 males and 9 females. The age range spanned from 26 to 52 years, yielding a mean age of 36.6 years with a standard deviation of 7.83. The participant pool represented varied academic and professional backgrounds, consisting of 8 students, 7 post-doctoral researchers, and 10 employees. Prior to conducting the experimental assessment, participants completed a comprehensive questionnaire designed to capture demographic information, including age, gender, and employment status. Additionally, the questionnaire gathered data regarding participants' environmental consciousness, video streaming habits, and content preferences to establish baseline behavioral patterns. To ensure methodological rigor and familiarize participants with the experimental variables, a training phase was implemented utilizing video sequences extracted from the Big Buck Bunny reference content. These training sequences were encoded at various resolutions and presented under different luminance conditions across the three device types. During these preparatory sessions, participants received detailed instructions regarding the quality assessment methodology, which employed the single discrete ACR scale. This standardized scale, conforming to ITU-R Rec. P.800 [82], comprised five categorical labels: Bad, Poor, Fair, Good, and Excellent. The experimental protocol required participants to evaluate the perceived QoE across all test conditions outlined in Table 4.1. For each device, the presentation sequence was structured to minimize visual fatigue, with the first eight videos displayed at minimum backlight luminance followed by a mandatory three-minute rest period. Subsequently, the remaining eight videos were presented at maximum backlight luminance, with other conditions randomized. This methodological decision was implemented to prevent potential visual discomfort from frequent backlight adjustments, which could adversely affect QoE ratings. Participant positioning was carefully controlled throughout the assessment. A sofa-based viewing arrangement was established, with viewing distances calibrated according to device specifications. For television viewing, distances adhered strictly to ITU-R BT.500 [84] guidelines. For laptop and smartphone assessments, participants were permitted to adjust viewing distances within comfortable parameters based on individual preferences. To maintain consistent experimental conditions, a three-minute rest period was enforced between device transitions. Each video presentation concluded with participants providing their QoE assessment using the aforementioned rating scale. The com-

Table 4.3: Results for the multi-way Analysis of Variance (ANOVA). $p < 0.05$ (*), $p < 0.01$ (**), $p < 0.001$ (***)

Var.	TV		LP		SP		ALL	
	F	p	F	p	F	p	F	p
BL	58.8	***	31.9	**	1200	***	341	***
AL	5.7	0.06	6.7	*	1.7	0.24	0.2	0.68
RES	0.1	0.76	45.7	**	0.2	0.68	12	*
CL	4.5	0.09	5.3	0.07	1.2	0.32	7.4	*

Table 4.4: T-test results between TCs with $BL = Min$ and TCs with $BL = Max$ while other variables are kept equal. $p < 0.05$ (*), $p < 0.01$ (**), $p < 0.001$ (***)

TCs compared	AL	Res	CL	TV	LP	SP	ALL
TC1 vs. TC9	0	H	H	***	**	***	***
TC2 vs. TC10	0	L	H	0.11	*	***	***
TC3 vs. TC11	0	H	L	**	*	***	***
TC4 vs. TC12	0	L	L	0.56	0.63	***	***
TC5 vs. TC13	500	H	H	***	**	***	***
TC6 vs. TC14	500	L	H	*	0.97	***	***
TC7 vs. TC15	500	H	L	**	0.07	***	***
TC8 vs. TC16	500	L	L	***	0.12	***	***

plete experimental protocol required approximately 30 minutes per participant to execute, ensuring comprehensive data collection while managing participant fatigue.

4.3 Results

Prior to computing the MOS and performing statistical analysis, outlier detection was conducted following the "final rejection criteria for discarding an observer of a test" specified in the ITU-R BT.500-14 [84]. Through this analysis, 4 outliers were identified for the TV data and 5 outliers each for the LP and SP data. After removing these outliers, the final valid dataset comprised responses from 21 participants for TV, 20 for LP, and 20 for SP. Notably, 23 out of the 25 total participants self-identified as environmentalists.

4.3.1 Mean Opinion Score Analysis

The MOS was computed for each of the 16 test conditions shown in Table 4.1, including 95% confidence intervals. Analyses were performed both for individual devices and averaged across all devices. When examining device backlight settings, configurations with minimum BL (TC1 through TC8) achieved an average MOS of

Table 4.5: T-test results between TCs with $RES = High$ and TCs with $RES = Low$ while other variables are kept equal. $p < 0.05$ (*), $p < 0.01$ (**), $p < 0.001$ (***)

TCs compared	BL	AL	CL	TV	LP	SP	ALL
TC1 vs. TC2	Min	0	H	*	0.07	0.48	0.92
TC3 vs. TC4	Min	0	L	0.64	0.27	0.12	0.76
TC5 vs. TC6	Min	500	H	0.19	0.61	0.49	0.49
TC7 vs. TC8	Min	500	L	0.06	*	0.37	0.09
TC9 vs. TC10	Max	0	H	0.85	**	0.83	0.09
TC11 vs. TC12	Max	0	L	**	**	0.72	***
TC13 vs. TC14	Max	500	H	0.85	**	0.69	0.06
TC15 vs. TC16	Max	500	L	0.42	**	0.57	*

2.83 across all devices - a full point lower than the average MOS of 3.89 obtained with maximum BL settings. The TV consistently achieved the highest MOS values across most test conditions, particularly excelling in scenarios with maximum BL. The LP demonstrated the second-highest MOS values under minimum BL conditions, with exceptions occurring in TC5 and TC7. Notably, video viewing on the SP with minimum BL received poor evaluations, yielding an average MOS of only 1.94. This notably low score likely stems from the SP's reduced minimum BL capability, implemented to optimize battery life. However, when maximum BL was employed, the SP achieved MOS values comparable to the LP for FHD videos and even surpassed them for HD videos, potentially due to reduced visibility of compression artifacts on the smaller SP display.

4.3.2 Analysis of Variance

Initial statistical evaluation employed one-way ANOVA to examine differences between ACR scores across the three devices. This analysis yielded an F-score of 57.35 with $p < 0.001$. Multiple comparison testing revealed statistically significant differences ($p < 0.001$) between mean ACR scores for all three devices, confirming that device type significantly impacts user QoE. Subsequently, multi-way ANOVA was conducted to analyze how multiple categorical variables influence mean outcomes. The analysis incorporated four categorical variables - BL, AL, CL, and Renewable Energy Source (RES) - with MOS serving as the outcome variable. Results presented in Table 4.3 confirm BL as a significant variable for MOS across all three devices. Additionally, AL and RES demonstrated statistical significance for the LP, while CL showed no significant impact for any individual device. However, when examining average MOS across all devices, BL, RES, and CL all emerged as statistically significant variables.

4.3.3 QoE versus Sustainability Trade-offs

This section examines the comparative analysis of MOS values achieved under different test conditions to identify potential trade-offs between acceptable QoE and environmental sustainability. The analysis focuses specifically on test conditions that exhibit two key characteristics: varying energy consumption levels and user-controllable parameters. The evaluation considers the backlight luminance and resolution as adjustable parameters, while maintaining constant values for ambient luminance and content luminance, which are beyond user control.

4.3.3.1 Impact of Backlight Luminance

The relationship between device power consumption and backlight luminance demonstrates a direct correlation, with magnitude varying by device type. As evidenced in Table 4.2, increasing backlight luminance leads to higher power consumption, with larger screens exhibiting greater sensitivity to these adjustments. This relationship suggests that reducing backlight luminance settings could yield substantial energy savings for end users. The impact of backlight luminance reduction on QoE varies significantly across device types. T-test analysis comparing minimum and maximum backlight luminance conditions, while controlling for other variables, reveals distinct patterns. For smartphone users, backlight luminance significantly influences QoE across all test conditions, suggesting that minimum backlight settings should only be employed when battery conservation is critical, given the substantial degradation in user experience. Television and laptop devices demonstrate more nuanced relationships between backlight luminance and QoE. For television viewing, backlight luminance significantly affects QoE under specific conditions: in dark ambient environments ($AL = 0$) when displaying 4K content, but shows no significant impact with FHD content. Conversely, in bright ambient conditions ($AL = 500$), backlight luminance significantly influences QoE regardless of resolution or content luminance. This finding suggests that users viewing FHD content in dark environments could reduce backlight settings to minimum levels, achieving up to 75% power reduction without meaningful QoE degradation. Laptop users experience significant backlight luminance effects on QoE in dark environments, except when viewing HD content with low content luminance. In bright environments, significant impacts only manifest when viewing FHD content with high content luminance. These results indicate numerous scenarios where users could reduce backlight settings without compromising QoE.

4.3.3.2 Impact of Resolution

Video resolution represents the most significant factor affecting end-to-end electricity consumption in video streaming, as discussed in Section 4.1.3 and demonstrated in Table 4.2. Reducing resolution, particularly from FHD or higher, offers substantial potential for energy and CO₂ emission reductions. However, the multi-way

ANOVA results suggest possible negative impacts on user QoE. T-test analysis comparing minimum and maximum resolution conditions reveals that resolution changes generally do not significantly impact QoE. This finding proves particularly robust for smartphone users, suggesting they could adopt HD rather than FHD content, achieving energy and emission reductions without meaningful QoE degradation. Television viewing shows significant MOS differences only under specific conditions in dark environments: with minimum backlight luminance and high content luminance, or maximum backlight luminance and low content luminance. These results suggest users can detect and appreciate higher resolutions only under optimal visual conditions, where content and display characteristics align favorably. Laptop users demonstrate significant resolution sensitivity when viewing content at maximum backlight luminance, regardless of ambient or content luminance. However, at minimum backlight settings, resolution differences become significant only in bright environments with low content luminance. These findings suggest that in most scenarios, users could select lower resolution content to achieve substantial energy savings while maintaining satisfactory QoE levels.

4.4 Outcomes and Discussion

This thesis has investigated the complex relationship between sustainable video streaming practices and user-perceived QoE, with particular focus on addressing the inherent tension between reducing energy consumption and CO_2 emissions while maintaining satisfactory service quality. Through a comprehensive subjective assessment methodology, we have examined the multifaceted impacts of video resolution, luminance configurations, and device characteristics on both QoE and energy consumption patterns in video streaming services. The subjective assessment results revealed significant insights into user experience across different device types. Television displays consistently achieved superior MOS values across the majority of test conditions, with particularly strong performance when configured with maximum backlight luminance. A notable finding emerged regarding the impact of backlight luminance settings - reducing to minimum backlight levels resulted in a substantial degradation of average MOS across all devices, decreasing by a full point from 3.89 to 2.83. This effect was especially pronounced for smartphone devices, where minimum backlight settings led to particularly poor user experience ratings with an average MOS of just 1.94. Through rigorous statistical analysis employing ANOVA and t-test methodologies, we identified distinct patterns of variable significance across device types. The analysis demonstrated that ambient luminance and resolution emerged as statistically significant variables exclusively for laptop devices, while content luminance showed no significant impact across individual device categories. However, when examining the aggregate MOS across all devices, backlight luminance, resolution, and content luminance all demonstrated statistical significance. The research findings suggest promising pathways toward sustainable video

streaming practices under specific viewing conditions. For instance, users watching content on smartphones' smaller screens, in low-light environments, or with reduced backlight settings could be encouraged to adopt more energy-efficient streaming parameters without significantly compromising their viewing experience. This is supported by our empirical evidence showing that increased energy expenditure through higher quality streaming does not necessarily translate to proportionally enhanced QoE under these conditions. Our findings lay the groundwork for future research directions focused on identifying and understanding additional factors that could motivate users to embrace sustainable streaming behaviors. The ultimate goal is to develop sophisticated QoE- and green-aware service management solutions that can effectively reduce CO_2 emissions while maintaining user satisfaction at acceptable levels. Such solutions would need to dynamically balance quality requirements with environmental impact, taking into account the various device-specific and environmental factors identified in this study. The research presented in this thesis contributes to the growing body of knowledge regarding sustainable multimedia services and provides crucial insights for developing next-generation video streaming platforms that can effectively balance environmental responsibility with user experience requirements.

The exploration of energy-aware resource management in 5G networks represents a crucial step toward sustainable network operations. The next chapter will delve deeper into the specific implementation details of this proposed solution in 5 with a contribution of the QoE-7 [17], including: The mathematical modeling of the resource allocation problem, the development of efficient algorithms for real-time decision making, and comprehensive performance evaluation metrics. This investigation will focus particularly on the practical aspects of deploying such solutions in real-world network environments, considering both technical and operational constraints. The subsequent chapter will present detailed experimental results from our implementation 4.1 of this proposed solution, demonstrating its effectiveness in achieving the desired balance between user experience and energy efficiency in modern communication networks. This analysis will provide valuable insights into the practical applicability of our approach and its potential impact on future network deployments.

Chapter 5

Towards Sustainable 5G: QoE-Driven Energy-Aware Resource Allocation

The rapid advancement of multimedia services and increasing network traffic have led to significant energy consumption challenges in modern communication networks. As a natural progression of this research, we propose an innovative solution for dynamic resource allocation in 5G heterogeneous network environments. This approach aims to establish an optimal balance between the QoE perceived by network users consuming video content and the overall network energy consumption taken the scenario from the previous Chapter 4.

The proposed solution addresses several critical aspects of modern network management. First, it considers the diverse requirements of different user devices and their impact on both quality perception and energy consumption combining the scenario which is described in the Chapter 4.1. Second, it incorporates adaptive resource allocation mechanisms that can dynamically respond to changing network conditions while maintaining energy efficiency. Third, it implements a comprehensive utility function that combines both QoE and energy metrics to guide resource allocation decisions with a contribution of the thesis QoE-7 [17].

The convergence of QoE management and sustainability in 5G networks presents unique challenges and opportunities. This section examines a novel approach to resource allocation that optimizes both user experience and energy efficiency across heterogeneous devices. The proposed framework demonstrates that substantial reductions in network energy consumption can be achieved while maintaining satisfactory QoE, particularly for mobile device users, while offering flexibility to television and laptop users based on their individual sustainability preferences. The evolution of mobile network technologies has contributed significantly to energy efficiency objectives. The 5G New Radio (NR) standard incorporates innovative

power management features, including sleep states for radio equipment that enable hardware components to be deactivated during periods of low traffic. Combined with reduced always-on signaling requirements, these capabilities facilitate extended sleep intervals during modest transmission loads, yielding meaningful reductions in network-wide energy consumption [125]. However, comprehensive understanding of mobile network energy dynamics remains incomplete. Conflicting evidence exists regarding the relative efficiency of 5G systems. While some research indicates that 5G networks achieve approximately 50% lower carbon emissions per bit transmitted compared to Fourth Generation Technology (4G) [126], other studies suggest that underutilized 5G bandwidth in burst transmission scenarios can result in 1.67 times greater energy consumption than 4G for equivalent web loading operations [127]. Independent of transmission infrastructure considerations, research by [114] reveals critical insights regarding the relationship between data traffic volumes, device characteristics, and energy consumption in video streaming applications. Their analysis demonstrates a direct correlation between data throughput and energy requirements, while also establishing that smaller form factor devices with reduced display resolutions achieve superior energy efficiency. Traditional network management approaches that prioritize maximum QoE delivery may therefore result in excessive energy consumption by allocating resources to support highest-quality content delivery, particularly high-resolution video streams that generate substantial data traffic. This represents a key limitation of existing literature on network management, where QoE-centric resource allocation algorithms typically disregard sustainability considerations [128–130]. Recent research suggests that energy consumption can be significantly reduced by targeting satisfactory rather than maximum QoE levels. Subjective assessment results from [16] (see in 4.3) demonstrate that quality reductions, particularly transitions from native to next-lower display resolutions, often produce minimal perceptible QoE impact while achieving substantial energy and resource savings. Building on these findings, this work presents a dynamic resource allocation solution for 5G Ultra-Dense Heterogeneous Network (HetNet) environments that optimizes the balance between aggregate user QoE and network energy consumption. The proposed framework incorporates differentiated device categories (television, laptop, and smartphone) with corresponding utility functions that integrate QoE and energy consumption impacts. Extensive simulations evaluate various QoE and energy priority weightings across two scenarios: homogeneous device populations and heterogeneous device mixtures. Results demonstrate that smartphone users consistently achieve optimal QoE with minimal energy consumption, while television users can reduce energy consumption by 50% with negligible QoE degradation. Laptop users retain flexibility to reduce energy consumption by approximately one-third through modest QoE trade-offs. In mixed device scenarios, overall network energy consumption can be reduced by nearly 50% without compromising QoE.

The remainder of this section is organized as follows: Section 5.1 presents the proposed system model, Section 5.2 details the solution methodology, Section 5.3 dis-

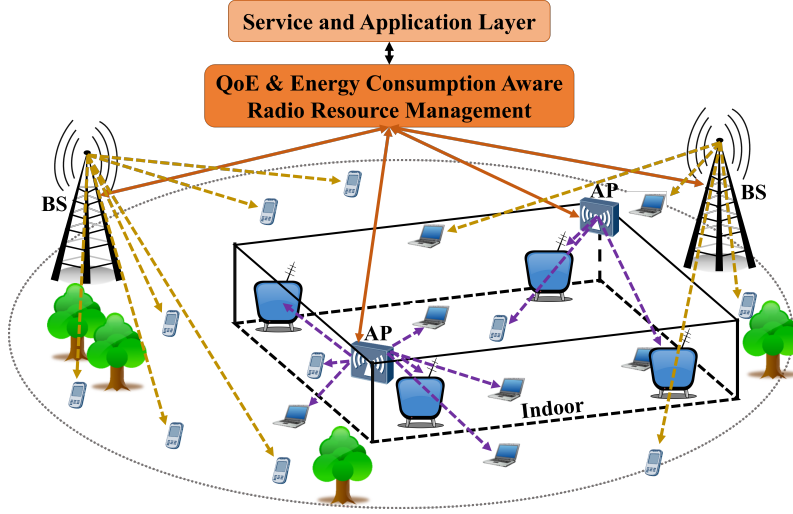


Figure 5.1: The considered HetNet deployment scenario.

cusses simulation results and performance analysis, and Section 5.4 provides concluding observations.

5.1 System Overview

This section presents a comprehensive system model designed to evaluate energy consumption and QoE in video content streaming through a 5G access network under varying network configurations. Figure 5.1 illustrates the proposed system architecture. The network infrastructure comprises a heterogeneous network HetNet deployment consisting of B NR base stations Base Station (BS)s defined by $\mathbb{B}$, where $b \in 1, 2, \dots, B$, alongside A NR access points Access Point (AP)s defined by $\mathbb{A}$, where $a \in 1, 2, \dots, A$. The system operates with a channel bandwidth W for both BSs and APs over an operating frequency f , encompassing N physical resource blocks Resource Block (RB)s. An RB, representing the smallest allocatable frequency resource for BSs and APs, consists of 12 consecutive and equally spaced subcarriers, yielding an RB bandwidth of $W_n = 12 \times \Delta f$. Following the 5G NR standard specifications, we implement a numerology (μ) defining the subcarrier spacing value as $\Delta f = 15 \times 2^\mu$ in kHz [131]. The network serves a set of D end-devices distributed randomly across the coverage area, denoted by $\mathbb{D}$ where $d \in 1, 2, \dots, D$. The system accommodates three distinct device categories: television (TV), laptop (LP), and smartphone (SP), represented as $i \in \text{TV}, \text{LP}, \text{SP}$. TVs and LPs maintain static positions with AP connectivity, while SPs exhibit random directional mobility with dynamic connectivity to either BSs or APs. Each device receives and displays video content at specific spatial resolutions, characterized by bitrate ($BR_{i,d}$). Resource allocation employs fast link adaptation procedures to manage available RBs at BSs and AP, selecting appropriate modulation coding schemes

Modulation Coding Schemes (MCS). This selection process occurs at each transmission time interval Transmission Time Interval (TTI) t' , incorporating channel quality information ($CQI_{i,d}$) from all devices, as detailed in [132].

For each device d , efficiency value $eff_{i,d}$ (bits/s/Hz) corresponds to channel reception conditions ($CQI_{i,d}$) [132]. The assigned bitrate $BR_{i,d}$ (bit/s) is calculated using Equation 5.1, where $n_{i,d}$ represents the number of RB allocated for video content reception at the specified resolution.

$$BR_{i,d} = n_{i,d} \times W_n \times eff_{i,d}, \quad (5.1)$$

Quality assessment incorporates both QoE and energy consumption metrics based on research presented in [16] (see in Chapter 4). The predicted Mean Opinion Score (MOS) for device d , denoted as $MOS_{i,d}^p$, is derived as a function of video bitrate $BR_{i,d}$ and device backlight luminance level ($BL_{i,d}$), expressed as $MOS_{i,d}^p = f(BR_{i,d}, BL_{i,d})$. Detailed QoE model specifications for each device type are elaborated in Section 5.2.1.

Energy consumption metrics encompass end-to-end evaluation from data center to end device. The overall electricity consumption $Q_{i,d}$ (kWh) follows Equation 5.2 [114], incorporating weekly streaming duration $t_{i,d}$, device power load $P_{i,d}$, data traffic $R_{i,d}$, and electricity intensity factor ρ .

$$Q_{i,d} = t_{i,d} \times (P_{i,d} + R_{i,d} \times \rho), \quad (5.2)$$

Resolution-dependent data traffic ranges from 1 GB/h for HD to 9 GB/h for 4K content, while power consumption varies significantly across device types, from several watts for SPs to over 200W for 4K TVs at maximum brightness. Comprehensive parameter values are documented in Chapter 4. System performance optimization balances video bitrate and device backlight luminance, where network management controls bitrate allocation while users determine backlight settings. The proposed framework implements centralized management (Figure 5.1) utilizing a utility function combining QoE and energy metrics with adjustable weighting factors. The resource management module processes device Channel Quality Indicator (CQI) data to determine optimal video bitrate assignments based on specified quality-energy trade-off preferences.

5.2 Proposed Solution

This section presents the development and implementation of our proposed solution for QoE-aware resource allocation in 5G networks. We first introduce the QoE modeling approach for different device types, followed by the resource allocation algorithm design.

5.2.1 QoE Modelling

The foundation of our solution lies in developing device-specific QoE models tailored to three representative end-user devices: television (60" UHD 4K screen, 3840×2160 pixels), laptop (15.6" FHD screen, 1920×1080 pixels), and smartphone (5.5" FHD screen, 1920×1080 pixels) (see in 4.1). These models derive from an extensive dataset of subjective quality assessments conducted in [16] (see in 4.3), which revealed the intricate relationships between device backlight luminance and video bitrate on both perceived quality and energy consumption in video streaming scenarios. The modeling process considered two primary independent variables - bitrate (BR) and backlight luminance (BL) - with the MOS serving as the dependent variable. While the original assessment in [16] included ambient luminance and content luminance as additional variables, these were excluded from our final models due to their negligible impact on QoE and energy consumption. Furthermore, these factors lie outside the realm of system control, being inherently dependent on environmental conditions and content characteristics.

We employed Linear Regression modeling techniques to establish the mathematical relationships between BR, BL, and MOS for each device category. The selection of linear regression models was justified by our dataset characteristics. To ensure robust validation of our models, we implemented 5-fold cross-validation procedures. The resulting device-specific QoE models are expressed through equations (5.3), (5.4), and (5.5), representing television, laptop, and smartphone devices respectively.

$$MOS_{TV}^p = 0.030 \times BR_{TV} + 0.708 \times BL_{TV} + 3.426 \quad (5.3)$$

$$MOS_{LP}^p = 0.606 \times BR_{LP} + 0.506 \times BL_{LP} + 2.797 \quad (5.4)$$

$$MOS_{SP}^p = -0.025 \times BR_{SP} + 1.975 \times BL_{SP} + 1.950 \quad (5.5)$$

The coefficient weights in these regression models reveal important insights about quality perception across different devices. For smartphone users, backlight luminance emerged as the dominant factor influencing QoE, with low backlight settings significantly degrading the viewing experience. Television and laptop users also demonstrated sensitivity to backlight variations, though to a lesser extent. Regarding bitrate impacts, laptop users showed the highest sensitivity to video quality changes, while television and smartphone users exhibited minimal quality perception differences at different bitrates. Notably, smartphone users occasionally rated lower resolution content more favorably, likely due to the limitations in discerning quality differences on smaller displays. To validate the models' predictive accuracy, we employed two key performance metrics: RMSE and PCC. The television model achieved an RMSE of 0.26 and PCC of 0.80, the laptop model demonstrated an RMSE of 0.21 and PCC of 0.87, while the smartphone model exhibited the highest accuracy with an RMSE of 0.12 and PCC of 0.99. These metrics represent averaged values across the 5-fold cross-validation process, indicating robust and reliable predictive capabilities across all device types.

5.2.2 QoE and Q based Resource Allocation Algorithm

Following the establishment of device-specific QoE models, we have developed a heuristic optimization solution for resource allocation within a 5G heterogeneous environment. The cornerstone of our decision-making framework is a dimensionless utility function ($UF_{i,d}$) that synthesizes both energy consumption ($Q_{i,d}$) and QoE ($MOS_{i,d}^p$) metrics for each device d . This function is mathematically expressed as:

$$UF_{i,d} = \alpha \times Q_{i,d}^N + (1 - \alpha) \times MOS_{i,d}^{p,N}, \quad (5.6)$$

where α serves as a weighting parameter to balance the relative importance of electricity consumption against user experience. Given that $Q_{i,d}$ and $MOS_{i,d}^p$ possess inherently different units of measurement, direct comparison would be mathematically incongruous [133]. Therefore, normalization is applied to ensure proportional contribution to $UF_{i,d}$, constraining its value within the interval $[0,1]$.

The normalized MOS ($MOS_{i,d}^{p,N}$) is derived through implementation of the utility function upward criterion [134], defined as:

$$MOS_{i,d}^{p,N} = \begin{cases} 0, & \text{s.t. } Cond_1 \\ 1 - \frac{MOS_i^{max} - MOS_{i,d}^p}{\delta' \times (MOS_i^{max} - MOS_i^{min})}, & \text{s.t. } Cond_2 \\ 1, & \text{otherwise.} \end{cases} \quad (5.7)$$

where:

$Cond_1$: $MOS_{i,d}^p < MOS_i^{min}$,

$Cond_2$: $MOS_i^{min} \leq MOS_{i,d}^p \leq MOS_i^{max}$. The parameters MOS_i^{min} and MOS_i^{max} represent the minimum and maximum MOS values corresponding to device type i , video bitrate $BR_{i,d}$, and backlight luminance level $BL_{i,d}$. The steepness parameter $\delta' \geq 2$ is tuned for optimization [133].

Conversely, the normalized energy consumption $Q_{i,d}^N$ is computed using the utility function downward criterion [134]:

$$Q_{i,d}^N = \begin{cases} 1, & \text{if } Q_{i,d} < Q_i^{min} \\ 1 - \frac{Q_{i,d} - Q_i^{min}}{\delta' \times (Q_i^{max} - Q_i^{min})}, & \text{if } Q_i^{min} \leq Q_{i,d} \leq Q_i^{max} \\ 0, & \text{otherwise,} \end{cases} \quad (5.8)$$

where Q_i^{min} and Q_i^{max} denote the minimum and maximum tolerated energy consumption values for device type i , $BR_{i,d}$, and $BL_{i,d}$, computed using equation (5.2). The system implements an iterative process at each transmission time interval t' to detect new network devices and optimize their association with either BS or AP to maximize $UF_{i,d}^{t'}$ ($UF_{i,d}^{max,t'}$). Furthermore, the algorithm re-executes upon detection of reception condition variations ($CQI_{i,d}^{t'}$ feedback) for any existing device d . This dynamic optimization may result in either network handover or updates to MCS and corresponding $n_{i,d}^{t'}$ within the current network. The selection and handover processes incorporate both network and user perspectives, utilizing QoE and energy

consumption as decisive criteria.

The algorithmic optimization can be formulated as a long-term utility maximization problem for $UF_{i,d}$, $\forall d \in \mathbb{D}$:

$$\max \sum_{t'} \sum_d UF_{i,d}^{t'} \quad (5.9a)$$

$$s.t. \ BR_{i,d}^{t'} \geq BR_i^{min}, \quad (5.9b)$$

where BR_i^{min} represents the minimum bitrate threshold for device type i .

5.3 Performance Evaluation

This section presents a comprehensive assessment and analysis of the proposed solution through extensive simulation experiments conducted in a recreated 5G heterogeneous environment for video service delivery to multiple types of end devices. The evaluation framework employs an ad-hoc 5G link-level simulator Link-Level Simulator (LLS) [135] implemented in Python. The following subsections detail the simulation parameters and configurations, followed by an in-depth analysis of the achieved results.

5.3.1 Simulation Settings

The simulation environment replicates a heterogeneous network architecture comprising one outdoor NR micro-BS ($B = 1$) and two indoor NR APs ($A = 2$). The system operates at a frequency $f = 28$ GHz, implementing an NR numerology $\mu = 3$ which yields a resource block bandwidth W_n of 1.44MHz, following specifications detailed in [131, 136]. The network is configured to support a maximum capacity of 100 connected devices ($D = 100$), each receiving and displaying video content. Table 5.1 provides a comprehensive overview of the primary simulation parameters. While these specific configurations are employed in our evaluation, it is important to note that the proposed system architecture maintains flexibility for adaptation to alternative network configurations.

The performance evaluation encompasses two distinct simulation scenarios:

- **Scenario A- Homogeneous Device Distribution:** In this configuration, all network devices (100%) are configured as a single device type - either TVs, LPs, or SPs. This scenario enables isolated analysis of device-specific behaviors and impacts concerning utility function considerations. Multiple BL configurations are evaluated to provide comprehensive insights into luminance effects on performance metrics.
- **Scenario B- Heterogeneous Device Distribution:** This scenario implements a more realistic device distribution, with 75% of connected devices configured

Table 5.1: Simulation parameters.

Parameter	Value
Number of BS/AP	1/2
BS/AP Operating frequency	28 GHz [131, 136]
BS/AP Bandwidth , W	400 MHz
NR numerology, μ	3
RB 's bandwidth, W_n	1.44 MHz
Subcarrier spacing	120 kHz
Transmission power AP/BS	40 dBm/26 dBm
Power Spectral density of noise	-174 dBm/Hz
Height BS/AP	10 m/4 m
Small/Large-scale fading models	[135]
Dynamic line of sight	Yes
Number of end devices	100
Type of end devices	TV, LP, SP
Height of end devices	1.5 m
Mobility Model TV/LP/SP	Static/Static/Random directional [135]
Reception mode TV/LP/SP	Indoor/Indoor-Outdoor/Indoor-Outdoor
SP' speed	0-0.8 m/s
Service type	Video
Video resolution TV/LP/SP	4K-FHD/FHD-HD/FHD-HD
Video bitrate (s.t., resolution) TV/LP/SP (Mbps)	20-5/5-2/5-2 [16]
Backlight luminance TV/LP/SP (lx)	5500-300/4000-400/5000-200 [16]

as SPs, while the remaining 25% are randomly distributed between TVs and LPs. The BL values are randomly assigned across devices. This configuration facilitates evaluation of utility function considerations in a mixed-device environment. To ensure statistical significance, results are averaged across 30 simulation runs, establishing a 95% confidence interval.

For clarity in result presentation, $BL = 0$ and $BL = 1$ in Figures 5.2, 5.3, and 5.4 denote the darkest and brightest BL values respectively for each device type, with actual luminance values specified in Table 5.1.

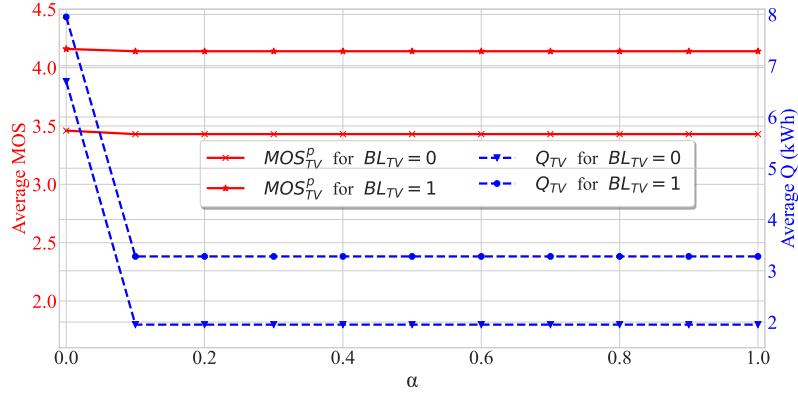


Figure 5.2: Average MOS_{TV}^p and Q_{TV} for UF_{TV}^{max} and different α values.

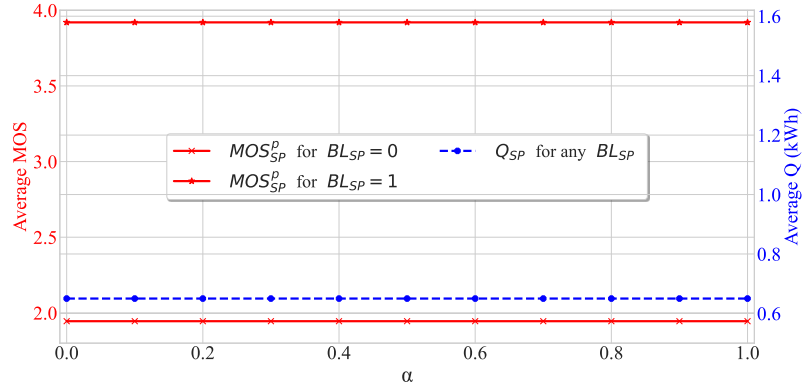


Figure 5.3: Average MOS_{SP}^p and Q_{SP} for UF_{SP}^{max} and different α values.

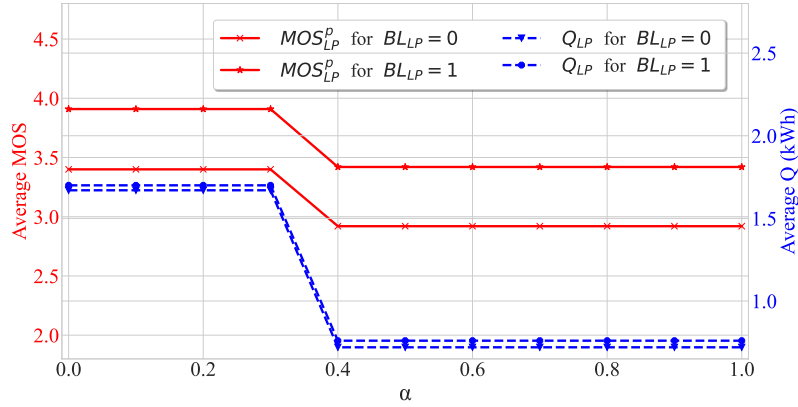


Figure 5.4: Average MOS_{LP}^p and Q_{LP} for UF_{LP}^{max} and different α values.

5.3.2 Results and Analysis

This section presents a comprehensive analysis of the simulation results for both considered scenarios, examining the relationships between various performance metrics

and system parameters. The analysis provides insights into the effectiveness of the proposed resource allocation approach in achieving an optimal balance between QoE and energy consumption.

5.3.2.1 Scenario A Results

The initial analysis focuses on Scenario A, where all network devices are of the same type, enabling a detailed examination of device-specific behaviors. The relationship between $MOS_{i,d}^p$ and $Q_{i,d}^p$ is analyzed across different α values, demonstrating how the utility function optimization affects performance metrics for TVs, SPs, and LPs independently. For the TV scenario, the results reveal that $MOS_{TV,d}^p$ is predominantly influenced by the backlight luminance, with bitrate having minimal impact due to the weight coefficient in equation (5.3). However, $BR_{TV,d}$ variations significantly affect $Q_{TV,d}$ due to the substantial difference in data traffic requirements between 4K and FHD video content. When $\alpha = 0$, the system prioritizes MOS, achieving $UF_{TV,d}^{max}$ with maximum bitrate, MOS, and energy consumption. For any positive α value, the optimization process reduces $BR_{TV,d}$, resulting in lower $Q_{TV,d}$ and $MOS_{TV,d}^p$ values. The reduction from 4K to FHD resolution produces negligible QoE degradation while halving electricity consumption. Additionally, adjusting the TV's backlight luminance from maximum to minimum results in a modest QoE reduction of 0.5 MOS points while decreasing energy consumption. In the SP scenario, bitrate's contribution to $MOS_{SP,d}^p$ remains minimal but exhibits a negative correlation. Even with $\alpha = 0$, the system achieves $UF_{SP,d}^{max}$ with minimum energy consumption ($Q_{SP,d}^{min}$), minimum bitrate ($BR_{SP,d}^{min}$), and maximum QoE ($MOS_{SP,d}^{max}$) across all backlight configurations. This behavior indicates that the optimization process inherently adopts an energy-efficient profile for smartphones. The results suggest that reducing backlight luminance on smartphones is counterproductive, as it significantly degrades QoE from good to poor without providing energy savings. For LP devices, both bitrate and backlight luminance demonstrate comparable impacts on $MOS_{LP,d}^p$, reflecting the balanced nature of the laptop QoE model. With $\alpha < 0.4$, the system achieves $UF_{LP,d}^{max}$ through maximum bitrate and QoE, despite increased energy consumption ($Q_{LP,d}^{max}$). Users prioritizing sustainability can accept a QoE reduction of 0.5 MOS points to achieve one-third reduction in energy consumption when $\alpha \geq 0.4$. This trade-off becomes more acceptable with maximum backlight settings, where the MOS decreases from 4.0 to 3.5, compared to minimum backlight settings where MOS drops from 3.5 to 3.0.

5.3.2.2 Scenario B Results

The examination of Scenario B provides insights into system performance with a heterogeneous device distribution. The average energy consumption ($Q_{i,d}$) analysis across varying network device populations demonstrates that setting $\alpha = 0$ leads to increased energy consumption due to MOS prioritization. However, the marginal

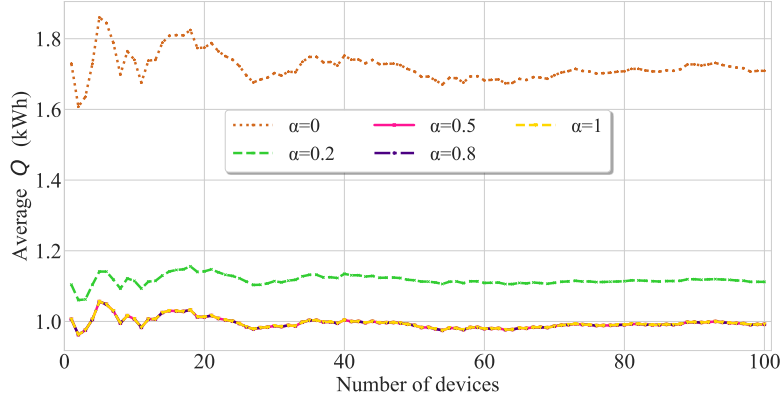
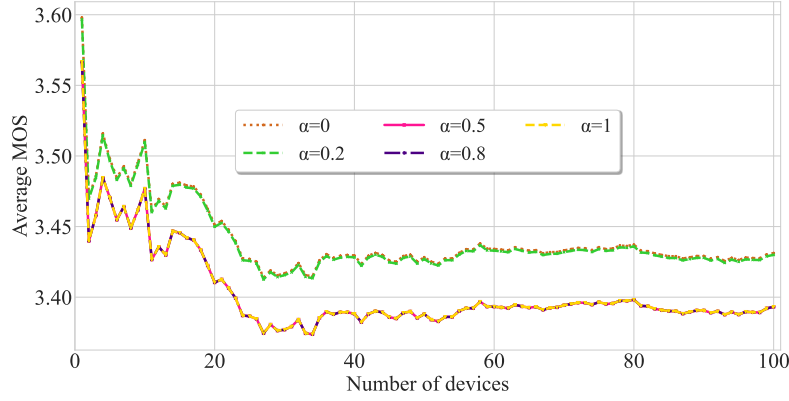
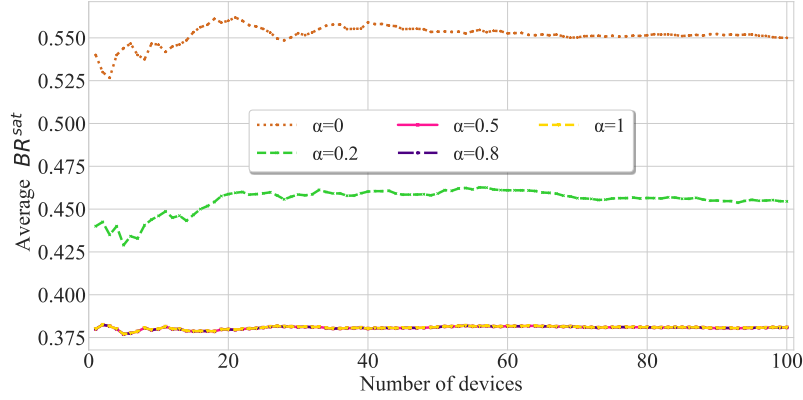


Figure 5.5: Average Q for different numbers of devices in the network.

QoE improvement between $\alpha = 0$ and $\alpha = 0.2$ does not justify the additional energy expenditure. Consistent with device-specific findings, optimization with $\alpha > 0$ for SPs and TVs, and $\alpha \geq 0.4$ for LPs, achieves energy minimization. Consequently, average $MOS_{i,d}^p$ values for $\alpha \in 0.5, 0.8, 1.0$ are lower than those observed with $\alpha \in 0, 0.2$, demonstrating the system's capability to achieve energy savings through modest QoE compromises. The analysis of bitrate satisfaction ($BR_{i,d}^{sat}$) reveals that even with $\alpha = 0$, the average satisfaction remains below 0.57. While this might appear suboptimal when considering bitrate alone, the holistic analysis demonstrates that reduced overall bitrate allocation achieves energy and resource savings without significantly impacting perceived quality. For instance, a TV with $CQI_{TV,d} = 9$ requires four resource blocks for maximum bitrate (20 Mbps) but only one resource block for minimum bitrate (5 Mbps). The results demonstrate that with $\alpha \geq 0.4$, the system optimizes $UF_{i,d}$ while minimizing $Q_{i,d}$ and resource block requirements. In the specific case of 100 network devices, implementing $\alpha \geq 0.4$ achieves a 22% reduction in required resource blocks compared to $\alpha = 0$. The findings confirm that bitrate variations minimally impact $MOS_{i,d}$ for SPs and TVs, while LP management requires careful α adjustment based on user expectations and sustainability preferences. The proposed allocation strategy thus provides a comprehensive solution balancing overall QoE with energy awareness.

5.4 Outcomes and Discussion

This thesis has presented a novel solution for dynamic resource allocation in 5G HetNet scenarios, addressing the critical balance between network infrastructure energy consumption and end-user QoE. The fundamental premise of our approach stems from the observation that increasing network resource allocation does not necessarily yield proportional improvements in perceived quality, while invariably driving higher energy consumption. To validate our proposed methodology, we con-

Figure 5.6: Average MOS for different numbers of devices in the network.Figure 5.7: Average BR^{sat} for different numbers of devices in the network.

structured a comprehensive simulation environment incorporating a 5G access network serving 100 devices across three distinct categories: television sets, laptops, and smartphones. For each device category, we developed specialized QoE and energy consumption models that account for two primary variables: video bitrate and device backlight luminance levels. The simulation framework employed a utility function with a configurable importance parameter (α) to evaluate the trade-offs between QoE and energy efficiency. Our analysis revealed distinct patterns when examining homogeneous device deployments. For television devices, reducing video bitrate resulted in minimal QoE degradation while achieving greater than 50% reduction in energy consumption. Smartphone configurations demonstrated consistent energy consumption patterns regardless of bitrate and backlight settings, though the darkest backlight settings significantly impaired QoE. This finding suggests smartphone users can maintain an environmentally conscious profile while utilizing maximum backlight settings. Laptop users faced a more nuanced decision matrix, with the option to sacrifice 0.5 MOS points in exchange for approximately 33% reduction in energy consumption. When examining heterogeneous device distributions

(75% smartphones, 25% televisions and laptops), our results demonstrated substantial potential for energy optimization with minimal QoE impact. Specifically, implementing $\alpha = 0.2$ achieved nearly 50% energy reduction without perceptible QoE degradation, while $\alpha = 0.5$ extended energy savings beyond 50% with only marginal QoE reduction (0.15 MOS points). While acknowledging that our findings are partially contingent on the specific subjective data and device types considered, the fundamental methodology remains extensible to diverse and updated QoE and energy consumption models. This adaptability is particularly relevant given the current paucity of comprehensive models in the literature [6-13]. Future research directions should prioritize deeper exploration of the relationship between QoE and energy consumption patterns. Such investigations will enable more precise tuning of network management approaches to optimize the balance between service sustainability and user satisfaction. The results presented in this thesis demonstrate the viability of implementing energy-aware resource allocation strategies in next-generation wireless networks while maintaining acceptable quality levels for diverse multimedia services. This work contributes to the broader goal of developing sustainable ICT infrastructure that satisfies both environmental imperatives and user experience requirements.

Chapter 6

Research Insights and Next Steps

6.1 Key Research Findings

This doctoral thesis has advanced real-time communications quality prediction through systematic investigation and innovative methodological integration. The research has progressed through several interconnected studies, each building upon previous findings to create a comprehensive framework for QoE estimation and sustainable service delivery.

In the contribution of the thesis QoE-1 (see in the Section 2.2), the initial investigation [10] established fundamental understanding through rigorous experimental analysis of network impairments' impact on user experience in WebRTC applications. Through controlled experiments involving structured video conversations, we quantified the complex relationships between network conditions and perceived quality while examining the differential sensitivity of audio and video components to various impairments. This foundational study revealed that audio QoE consistently demonstrated superior performance compared to video QoE metrics, providing crucial insights for optimizing real-time communication systems. The research also uncovered significant asymmetry in the impact of network impairments across different media components, with implications for adaptive quality management strategies.

After, the investigation expanded to application-layer analysis with a contribution of the thesis QoE-2 (see in the Section 2.3) [11], focusing on WebRTC session parameters and their correlation with user experience. Through comprehensive statistical validation employing ANOVA and PCC analysis, we identified key performance indicators enabling the development of sophisticated machine learning models for QoE estimation. This phase demonstrated the effectiveness of our statistical analysis process in identifying optimal application-level parameters, with the MLP architecture achieving state-of-the-art results ($R^2 = 0.852$, RMSE = 0.282).

A significant advancement was achieved through facial expression analysis with a contribution of the thesis QoE-3 (see in the Section 2.4) [12], implementing both traditional machine learning approaches and advanced FCN architectures. This phase

developed novel preprocessing methods for facial AUs analysis, revealing optimal feature selection strategies and model architectures. The research demonstrated superior performance of preprocessed feature selection (0.78 accuracy) compared to raw feature processing through FCN (0.70 accuracy), establishing important benchmarks for facial-based QoE estimation.

In the QoE-4 (see in the Section 2.5) [13], the research reached a crucial milestone through the integration of facial and speech features [13] using sophisticated ICKA data fusion techniques. This multimodal approach achieved remarkable accuracy (0.93), substantially outperforming single-modality approaches (0.78 for facial and 0.86 for speech features). The success of this integration demonstrated the value of comprehensive feature analysis, incorporating 129 statistically significant features (111 voice-related and 18 facial-related) to capture nuanced aspects of user experience.

The investigation addressed critical sustainability challenges in video streaming services with a contribution of the QoE-5 (see in the Chapter 3) [15] after investigating the WebRTC service, identifying the essential balance between energy consumption and user satisfaction. Through extensive subjective assessments and real-world experiment with the contribution of the QoE-6 (see in the Chapter 4) [16], we examined the complex interactions between video resolution, luminance configurations, and device characteristics impact on both QoE and energy consumption patterns. This research revealed opportunities for significant energy optimization while maintaining acceptable quality levels, particularly for mobile device users. The culmination of our research resulted in the development and validation of a dynamic resource allocation solution for 5G HetNet scenarios with a contribution of the thesis QoE-7 (see in the Chapter 5) [17]. This comprehensive solution demonstrated that network energy consumption could be reduced by more than 50% while maintaining satisfactory QoE levels across diverse device types. The findings showed particularly promising results for smartphone users, while providing television and laptop users flexibility in balancing quality and sustainability preferences.

Future work should explore bringing these research results to standardization bodies and terminal manufacturers. Our QoE estimation based on facial and speech features could inform WebRTC implementations, while our findings on energy-aware solutions can be considered in future video communication applications. Terminal manufacturers could benefit from understanding the relationship between device settings (like backlight luminance) and both QoE and energy consumption, potentially incorporating these insights into future product designs. Future research directions cover several critical areas:

- The extension of our multimodal methodology to broader application scenarios represents a primary research direction, investigating potential application-independent correlations between identified features and QoE. This includes developing generalized models incorporating application-specific parameters while maintaining high estimation accuracy across diverse use cases.

- Advanced machine learning approaches, particularly Multi-view learning techniques, warrant investigation for improving generalization efficiency across varying parameter sets. This direction aims to address the challenges of heterogeneous data sources and varying application configurations while maintaining robust performance.
- Deeper investigation of QoE-energy consumption relationships is essential for optimizing network management strategies. This includes developing more sophisticated models of user behavior and device energy consumption patterns, enabling more precise trade-offs between quality and sustainability.
- Further research into sustainable ICT infrastructure development is needed, focusing on holistic approaches that consider both environmental impact and user experience requirements. This includes investigating novel techniques for energy-aware service delivery and developing more comprehensive metrics for sustainability assessment.
- Finally, the integration of emerging technologies and standards into our framework represents an ongoing research challenge. This includes adapting our methodologies to new communication protocols, codec technologies, and device capabilities as they evolve.

In conclusion, the thesis has established an innovative foundation for multimodal QoE assessment in real-time communications while addressing critical sustainability challenges. Our findings have advanced the understanding of human-centric quality prediction through the innovative combination of network analysis, machine learning, and multimodal feature integration. The developed frameworks and methodologies provide essential building blocks for next-generation communication systems that can deliver high-quality services while minimizing environmental impact. The comprehensive insights derived from this research contribute to both theoretical understanding and practical application in the field of real-time communication systems. The study has demonstrated the effectiveness of integrated approaches to quality assessment and resource management, while highlighting important directions for future research in this rapidly evolving domain.

6.2 Research Challenges and Solutions

Throughout this research, several significant challenges were encountered and addressed. Collecting synchronized facial and speech data during natural communication posed significant methodological challenges. The need to capture spontaneous reactions while maintaining experimental control required careful balance. These challenges were addressed through careful experimental design with controlled network impairment introduction and high-quality recording equipment. The developed methodology incorporated precise synchronization between network condition

changes and multimodal data recording, ensuring temporal alignment between quality variations and user responses. The extraction of thousands of potential features created computational challenges for statistical analysis. The initial feature space included 6,373 speech features alone, necessitating efficient dimensionality reduction approaches. This challenge was resolved through systematic feature selection using ANOVA with stringent significance thresholds. By applying a threshold of $p < 0.01$ for speech features, we identified the most relevant indicators while maintaining computational tractability. The natural tendency of participants to avoid extreme ratings created imbalanced datasets that complicated machine learning model training. Class imbalance was particularly pronounced for the highest and lowest quality ratings, potentially biasing models toward middle-range predictions. The implementation of the ADASYN algorithm successfully mitigated this issue while preserving data distribution characteristics. Comparative analysis with other augmentation techniques confirmed ADASYN's superior performance in maintaining feature distribution properties. Finding optimal strategies for combining facial and speech features proved challenging given their different temporal characteristics. Speech features typically exhibit higher temporal variability, while facial expressions may persist across longer timeframes. Comparative analysis of multiple fusion techniques led to the selection of ICKA for its superior performance. The selected approach effectively accommodated the different statistical properties of the two modalities while maximizing their complementary information content. Accurately measuring device-specific energy consumption across different operational conditions required specialized instrumentation and methodology. Variations in power draw related to processing load, display brightness, and network activity needed to be isolated and quantified. This challenge was addressed using the Smart Plug TP-Link HS110 with calibrated measurement protocols. The measurement system incorporated statistical validation to account for temporal variations in power consumption, ensuring representative energy metrics across experimental conditions.

6.3 Scope and Limitations

While this research makes contributions, several limitations define its scope and indicate areas for future investigation.

The developed QoE estimation models are specifically validated for WebRTC-based audio-visual conversations and may require a reconfiguration of other communication protocols or streaming scenarios. The bidirectional nature of WebRTC communications creates specific patterns of network impairment and user response that may not generalize to unidirectional streaming or other multimedia applications. Future research should examine the transferability of these models to alternative communication contexts.

The participant pool for subjective assessment was limited to European users aged 23-36, potentially affecting the generalizability of emotional expression pat-

terns across different cultural contexts and age groups. Cultural differences in emotional expression and technology expectations may influence the relationship between observed behaviors and perceived quality. Broader demographic sampling would enhance model robustness across diverse user populations.

The energy consumption models were developed for specific device models and may require adjustment for different manufacturers or newer device generations. Technological advancements in display efficiency, processor architecture, and power management continue to evolve, potentially altering the energy-quality relationships identified in this research. Longitudinal studies tracking these relationships across device generations would provide valuable insights into technological trends.

The subjective assessment used a limited range of types of video content, which could underestimate the impact of content characteristics on QoE-energy relationships. Content factors such as genre, motion complexity, and semantic importance may influence quality perception and acceptable energy trade-offs. Expanded content diversity in future studies would enhance ecological validity and potentially reveal content-dependent optimization opportunities. The 5G network simulations focused on specific configurations and may not capture the full complexity of heterogeneous network deployments in real-world scenarios. Factors such as interference patterns, user mobility, and dynamic load variations introduce additional complexity that may affect resource allocation effectiveness. Field trials in operational networks would provide valuable validation of the simulation-based findings presented in this thesis. These limitations, while constraining the immediate application scope, also highlight promising directions for future research to extend and refine the methodologies and models developed in this thesis.

Bibliography

- [1] B. Garcia et al. “WebRTC testing: challenges and practical solutions”. In: *IEEE Communications Standards Magazine* 1.2 (2017), pp. 36–42.
- [2] X. Sandvine. “2023 global internet phenomena report”. In: (2023).
- [3] P. Le Callet et al. “Qualinet white paper on definitions of quality of experience”. PhD thesis. Qualinet (www. qualinet. eu), 2013.
- [4] K. De Moor et al. “Chamber QoE: a multi-instrumental approach to explore affective aspects in relation to quality of experience”. In: *Human Vision and Electronic Imaging XIX*. Vol. 9014. SPIE. 2014, pp. 204–217.
- [5] U. Engelke et al. “Psychophysiology-based QoE assessment: A survey”. In: *IEEE Journal of Selected Topics in Signal Processing* 11.1 (2016), pp. 6–21.
- [6] A. Bhattacharya, W. Wu, and Z. Yang. “Quality of experience evaluation of voice communication: an affect-based approach”. In: *Human-centric Computing and Information Sciences* 2 (2012), pp. 1–18.
- [7] S. Porcu et al. “Estimation of the quality of experience during video streaming from facial expression and gaze direction”. In: *IEEE Transactions on Network and Service Management* 17.4 (2020), pp. 2702–2716.
- [8] L. Itu-T. “Greenhouse gas emissions trajectories for the information and communication technology sector compatible with the unfccc paris agreement”. In: *document itu-t 1. 1470 (01/20)* (2020).
- [9] T. Hoßfeld et al. “A Greener Experience: Trade-Offs between QoE and CO₂ Emissions in Today’s and 6G Networks”. In: *IEEE communications magazine* 61.9 (2023), pp. 178–184.
- [10] G. Bingol et al. “The Impact of Network Impairments on the QoE of WebRTC applications: A Subjective study”. In: *2022 14th International Conference on Quality of Multimedia Experience (QoMEX)*. 2022, pp. 1–6.
- [11] M. Hamidi et al. “Analysis of Application-layer Data to Estimate the QoE of WebRTC-based Audiovisual Conversations”. In: *2023 IEEE Globecom Workshops (GC Wkshps)*. 2023, pp. 365–370.

- [12] G. Bingöl et al. “QoE Estimation of WebRTC-based Audiovisual Conversations from Facial Expressions”. In: *2022 16th International Conference on Signal-Image Technology Internet-Based Systems (SITIS)*. 2022, pp. 577–584.
- [13] G. Bingöl et al. “QoE Estimation of WebRTC-based Audio-visual Conversations from Facial and Speech Features”. In: *ACM Trans. Multimedia Comput. Commun. Appl.* 20.5 (2024).
- [14] G. Bingöl et al. “WebRTC-QoE: A dataset of QoE assessment of subjective scores, network impairments, and facial speech features”. In: *Computer Networks* 244 (2024), p. 110356.
- [15] G. Bingöl et al. “An Analysis of the Trade-Off Between Sustainability and Quality of Experience for Video Streaming”. In: *2023 IEEE International Conference on Communications Workshops (ICC Workshops)*. 2023, pp. 1600–1605.
- [16] G. Bingöl et al. “Are Quality and Sustainability Reconcilable? A Subjective Study on Video QoE, Luminance and Resolution”. In: *2023 15th International Conference on Quality of Multimedia Experience (QoMEX)*. 2023, pp. 19–24.
- [17] C. C. González et al. “A QoE-based Energy-aware Resource Allocation Solution for 5G Heterogeneous Networks”. In: *2024 16th International Conference on Quality of Multimedia Experience (QoMEX)*. 2024, pp. 29–35.
- [18] G. Bingöl et al. *WebRTC-QoE: A Dataset of Quality of Experience in Audio-Video Communications*. 2023.
- [19] S. Loreto and S. P. Romano. “How far are we from webrtc-1.0? an update on standards and a look at what’s next”. In: *IEEE Communications Magazine* 55.7 (2017), pp. 200–207.
- [20] H. Mahmoud and R. Abozariba. “A systematic review on WebRTC for potential applications and challenges beyond audio video streaming”. In: *Multimedia Tools and Applications* (2024), pp. 1–38.
- [21] H. Alvestrand. “RFC 8825: Overview: Real-Time Protocols for Browser-Based Applications”. In: (2021).
- [22] A. Gouaillard and L. Roux. “Real-time communication testing evolution with WebRTC 1.0”. In: *2017 Principles, Systems and Applications of IP Telecommunications (IPTComm)*. IEEE. 2017, pp. 1–8.
- [23] P. Le Callet, S. Möller, and A. Perkis. *Qualinet White Paper on Definitions of Quality of Experience (2012)*. European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003), Lausanne, Switzerland, Version 1.2, March 2013. 2012.

- [24] M. Schmitt et al. “Towards Individual QoE for Multiparty Videoconferencing”. In: *IEEE Transactions on Multimedia* 20.7 (2018), pp. 1781–1795.
- [25] L. F. Tiotsop et al. “Regularized Maximum Likelihood Estimation of the Subjective Quality from Noisy Individual Ratings”. In: *2022 14th Int. Conf. on Quality of Multimedia Experience (QoMEX)*. 2022, pp. 1–4.
- [26] C. Wang, H. Zhu, and M. Li. “SpeechQoE: A Novel Personalized QoE Assessment Model for Voice Services via Speech Sensing”. In: *Proc. of the 20th ACM Conf. on Embedded Networked Sensor Systems*. ACM, 2023, pp. 305–319.
- [27] M. J. Scott et al. “Modelling Human Factors in Perceptual Multimedia Quality: On The Role of Personality and Culture”. In: *Proc. of the 23rd ACM Int. Conf. on Multimedia*. ACM, 2015, pp. 481–490.
- [28] H. Yan et al. “Discovering Usage Patterns of Mobile Video Service in the Cellular Networks”. In: *IEEE Transactions on Network and Service Management* 18.2 (2021), pp. 1789–1802.
- [29] X. Liu et al. “An EEG-Based Study on Perception of Video Distortion Under Various Content Motion Conditions”. In: *IEEE Transactions on Multimedia* 22.4 (2020), pp. 949–960.
- [30] D. Egan et al. “An evaluation of Heart Rate and ElectroDermal Activity as an objective QoE evaluation method for immersive virtual reality environments”. In: *2016 Eighth Int. Conf. on Quality of Multimedia Experience (QoMEX)*. 2016, pp. 1–6.
- [31] U. Engelke et al. “Psychophysiology-Based QoE Assessment: A Survey”. In: *IEEE Journal of Selected Topics in Signal Processing* 11.1 (2017), pp. 6–21.
- [32] S. Li and W. Deng. “Deep Facial Expression Recognition: A Survey”. In: *IEEE Transactions on Affective Computing* 13.3 (2022), pp. 1195–1215.
- [33] R. A. Khalil et al. “Speech Emotion Recognition Using Deep Learning Techniques: A Review”. In: *IEEE Access* 7 (2019), pp. 117327–117345.
- [34] L. Skorin-Kapov et al. “A Survey of Emerging Concepts and Challenges for QoE Management of Multimedia Services”. In: *ACM Trans. Multimedia Comput. Commun. Appl.* 14.2s (2018).
- [35] N. Khan and M. Martini. “QoE-driven multi-user scheduling and rate adaptation with reduced cross-layer signaling for scalable video streaming over LTE wireless systems”. In: (2016).
- [36] Y. Chen, K. Wu, and Q. Zhang. “From QoS to QoE: A Tutorial on Video Quality Assessment”. In: *IEEE Communications Surveys & Tutorials* 17.2 (2015), pp. 1126–1165.

- [37] M. Seufert et al. “A Survey on Quality of Experience of HTTP Adaptive Streaming”. In: *IEEE Communications Surveys & Tutorials* 17.1 (2015), pp. 469–492.
- [38] F. Chiariotti. “A survey on 360-degree video: Coding, quality of experience and streaming”. In: *Computer Communications* 177 (2021), pp. 133–155.
- [39] S. Jelassi et al. “Quality of Experience of VoIP Service: A Survey of Assessment Approaches and Open Issues”. In: *IEEE Communications Surveys Tutorials* 14.2 (2012), pp. 491–513.
- [40] S. Barakovic and L. Skorin-Kapov. “Survey of research on Quality of Experience modelling for web browsing”. In: *Quality and User Experience* 6 (2017).
- [41] S. Vlahovic, M. Suznjetic, and L. Skorin-Kapov. “A survey of challenges and methods for Quality of Experience assessment of interactive VR applications”. In: *Journal on Multimodal User Interfaces* (2022), pp. 1–35.
- [42] B. Belmudez et al. “Audio and video channel impact on perceived audiovisual quality in different interactive contexts”. In: *2009 IEEE International Workshop on Multimedia Signal Processing*. IEEE. 2009, pp. 1–5.
- [43] I. RECOMMENDATION. “ITU-R P. 920-Interactive test methods for audiovisual communications”. In: (2000).
- [44] B. Belmudez and S. Möller. “Audiovisual quality integration for interactive communications”. In: *EURASIP Journal on Audio, Speech, and Music Processing* 2013 (2013), pp. 1–23.
- [45] F. Brauer, M. S. Ehsan, and G. Kubin. “Subjective evaluation of conversational multimedia quality in IP networks”. In: *2008 IEEE 10th Workshop on Multimedia Signal Processing*. IEEE. 2008, pp. 872–876.
- [46] B. Ristic et al. “A Tutorial on Bernoulli Filters: Theory, Implementation and Applications”. In: *IEEE Transactions on Signal Processing* 61.13 (2013), pp. 3406–3430.
- [47] K. De Moor et al. “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.
- [48] C. Tsiraras, M. Röscher, and B. Stiller. “VoIP-based Calibration of the DQX Model”. In: *2015 IFIP Networking Conference (IFIP Networking)*. IEEE. 2015, pp. 1–9.
- [49] ITU. “The E-model: a computational model for use in transmission planning”. In: Recommendation ITU-T G.107. 2015.
- [50] B. García et al. “Assessment of QoE for Video and Audio in WebRTC Applications Using Full-Reference Models”. In: *Electronics* 9.3 (2020).

- [51] L. Serreli et al. “Robust QUIC-Based Signalling for WebRTC in Impaired Networks”. In: *2023 IEEE International Mediterranean Conference on Communications and Networking (MeditCom)*. 2023, pp. 187–192.
- [52] S. Yan et al. “Predicting Freezing of WebRTC Videos in WiFi Networks”. In: *Ad Hoc Networks*. 2019, pp. 292–301.
- [53] D. Vučić and L. Skorin-Kapov. “The Impact of Packet Loss and Google Congestion Control on QoE for WebRTC-Based Mobile Multiparty Audiovisual Telemeetings”. In: *MultiMedia Modeling*. Ed. by I. Kompatsiaris et al. Cham: Springer International Publishing, 2019, pp. 459–470.
- [54] D. Vučić, L. Skorin-Kapov, and M. Sužnjević. “The impact of bandwidth limitations and video resolution size on QoE for WebRTC-based mobile multiparty video conferencing”. In: *Proc. 5th ISCA/DEGA Workshop on Perceptual Quality of Systems (PQS 2016)*. 2016, pp. 59–63.
- [55] D. Vučić and L. Skorin-Kapov. “QoE Assessment of Mobile Multiparty Audiovisual Telemeetings”. In: *IEEE Access* 8 (2020), pp. 107669–107684.
- [56] D. Ammar et al. “Exploring the Usefulness of Machine Learning in the Context of WebRTC Performance Estimation”. In: *2019 IEEE 44th Conference on Local Computer Networks (LCN)*. 2019, pp. 406–413.
- [57] D. Ammar et al. “Video QoE killer and performance statistics in WebRTC-based video communication”. In: *2016 IEEE Sixth International Conference on Communications and Electronics (ICCE)*. 2016, pp. 429–436.
- [58] J. B. Husić et al. “QoE Prediction of WebRTC Video Calls Using Google Chrome Statistics”. In: *2021 20th Int. Symposium INFOTEH-JAHORINA (INFOTEH)*. 2021, pp. 1–6.
- [59] P. Ekman and W. Friesen. “Constants across cultures in the face and emotion.” In: *Journal of personality and social psychology* 17.2 (1971), pp. 124–129.
- [60] P. Ekman and W. V. Friesen. “Facial action coding system: a technique for the measurement of facial movement”. In: 1978.
- [61] S. Afshari and N. Movahhedinia. “QoE assessment of interactive applications in computer networks”. In: *Multimedia Tools and Applications* 75.2 (2016), pp. 903–918.
- [62] L. Amour et al. “An improved QoE estimation method based on QoS and affective computing”. In: *2018 Int. Symposium on Programming and Systems (ISPS)*. 2018, pp. 1–6.
- [63] A. Bhattacharya, W. Wu, and Z. Yang. “Quality of experience evaluation of voice communication: an affect-based approach”. In: *Human-centric Computing and Information Sciences* 2.1 (2012).

- [64] S. Porcu, A. Floris, and L. Atzori. “Analysis of the quality of remote working experience: a speech-based approach”. In: *Quality and User Experience* 7.1 (2022).
- [65] J. Rößler, J. Sun, and P. Gloor. “Reducing Videoconferencing Fatigue through Facial Emotion Recognition”. In: *Future Internet* 13.5 (2021).
- [66] S. G. Koolagudi and K. S. Rao. “Emotion recognition from speech: a review”. In: *Int. Journal of Speech Technology* 15.2 (2012), pp. 99–117.
- [67] M. B. Mustafa et al. “Speech emotion recognition research: an analysis of research focus”. In: *Int. Journal of Speech Technology* 21.1 (2018), pp. 137–156.
- [68] P. Casas et al. “Next to You: Monitoring Quality of Experience in Cellular Networks From the End-Devices”. In: *IEEE Transactions on Network and Service Management* 13.2 (2016), pp. 181–196.
- [69] S. Kumar et al. “Client-Side QoE Management for SVC Video Streaming: An FSM Supported Design Approach”. In: *IEEE Transactions on Network and Service Management* 16.3 (2019), pp. 1113–1126.
- [70] J. Baraković Husić, S. Baraković, and A. Veispahić. “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)*. 2017, pp. 428–433.
- [71] B. Garcia et al. “Understanding and estimating quality of experience in WebRTC applications”. In: *Computing* 101.11 (2019), pp. 1585–1607.
- [72] ITU. “One-way transmission time”. In: Rec. ITU-T G.114. 2003.
- [73] D. Tsolkas et al. “A survey on parametric QoE estimation for popular services”. In: *Journal of Network and Computer Applications* 77.1 (2017), pp. 1–17.
- [74] G. Carlucci et al. “Analysis and Design of the Google Congestion Control for Web Real-Time Communication (WebRTC)”. In: *Proceedings of the 7th International Conference on Multimedia Systems*. MMSys ’16. Klagenfurt, Austria, 2016.
- [75] L. Rizzo. “Dummynet: a simple approach to the evaluation of network protocols”. In: *ACM SIGCOMM Computer Communication Review* 27.1 (1997), pp. 31–41.
- [76] S. Hemminger. “Network emulation with NetEm”. In: 2005.
- [77] N. Ahmad, A. Wahab, and J. Schormans. “Importance of Cross-Correlation of QoS metrics in Network Emulators to Evaluate QoE of Video Streaming Applications”. In: *2020 11th International Conference on Network of the Future (NoF)*. 2020, pp. 43–47.

- [78] I. Homoliak et al. “Exploitation of netem utility for non-payload-based obfuscation techniques improving network anomaly detection”. In: *International Conference on Security and Privacy in Communication Systems*. Springer. 2016, pp. 770–773.
- [79] V. Gupta and V. K. Sharma. “Skin typing: Fitzpatrick grading and others”. In: *Clinics in Dermatology* 37.5 (2019), pp. 430–436.
- [80] ITU. “Interactive test methods for audiovisual communications”. In: Recommendation ITU-T P.920. 2000.
- [81] ITU. “Effect of delays on telemeeting quality”. In: Recommendation ITU-T P.1305. 2016.
- [82] ITU. “Methods for subjective determination of transmission quality”. In: Recommendation ITU-T P.800. 1996.
- [83] M. M. Bradley and P. J. Lang. “Measuring emotion: The self-assessment manikin and the semantic differential”. In: *Journal of Behavior Therapy and Experimental Psychiatry* 25.1 (1994), pp. 49–59.
- [84] ITU. “Methodologies for the subjective assessment of the quality of television images”. In: Recommendation ITU-R BT.500-14. 2019.
- [85] B. Jansen et al. “Performance Evaluation of WebRTC-Based Video Conferencing”. In: *SIGMETRICS Perform. Eval. Rev.* 45.3 (2018), pp. 56–68.
- [86] T. Baltrusaitis et al. “OpenFace 2.0: Facial Behavior Analysis Toolkit”. In: *2018 13th IEEE Int. Conf. on Automatic Face Gesture Recognition (FG 2018)*. 2018, pp. 59–66.
- [87] E. Wood et al. “Rendering of Eyes for Eye-Shape Registration and Gaze Estimation”. In: *2015 IEEE Int. Conf. on Computer Vision (ICCV)*. 2015, pp. 3756–3764.
- [88] T. Baltrusaitis, M. Mahmoud, and P. Robinson. “Cross-dataset learning and person-specific normalisation for automatic Action Unit detection”. In: *2015 11th IEEE Int. Conf. and Workshops on Automatic Face and Gesture Recognition (FG)*. Vol. 06. 2015, pp. 1–6.
- [89] H. He et al. “ADASYN: Adaptive synthetic sampling approach for imbalanced learning”. In: *2008 IEEE Int. Joint Conf. on Neural Networks (IEEE World Congress on Computational Intelligence)*. 2008, pp. 1322–1328.
- [90] J. Long, E. Shelhamer, and T. Darrell. “Fully convolutional networks for semantic segmentation”. In: *2015 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2015, pp. 3431–3440.
- [91] T. Sainburg. *timsainb/noisereduce: v1.0*. Version db94fe2. 2019.
- [92] T. Sainburg, M. Thielk, and T. Q. Gentner. “Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires”. In: *PLoS computational biology* 16.10 (2020), e1008228.

- [93] B. McFee et al. “librosa: Audio and Music Signal Analysis in Python”. In: *Proc. of the 14th Python in Science Conf.* 2015, pp. 18–25.
- [94] F. Eyben, M. Wöllmer, and B. Schuller. “Opensmile: The Munich Versatile and Fast Open-Source Audio Feature Extractor”. In: *Proc. of the 18th ACM Int. Conf. on Multimedia.* ACM, 2010, pp. 1459–1462.
- [95] F. Eyben et al. “Recent Developments in OpenSMILE, the Munich Open-Source Multimedia Feature Extractor”. In: *Proc. of the 21st ACM Int. Conf. on Multimedia.* ACM, 2013, pp. 835–838.
- [96] B. Schuller et al. “The INTERSPEECH 2016 Computational Paralinguistics Challenge: Deception, Sincerity Native Language”. In: *Interspeech 2016.* 2016, pp. 2001–2005.
- [97] F. Weninger et al. “On the Acoustics of Emotion in Audio: What Speech, Music, and Sound have in Common”. In: *Frontiers in psychology* 4 (2013).
- [98] W. Liang et al. “Adaptive multiple kernel fusion model using spatial-statistical information for high resolution SAR image classification”. In: *Neurocomputing* 492 (2022), pp. 382–395.
- [99] L. Wang. *Support vector machines: theory and applications*. Vol. 177. Springer Science & Business Media, 2005.
- [100] P. Charonyktakis et al. “On User-Centric Modular QoE Prediction for VoIP Based on Machine-Learning Algorithms”. In: *IEEE Transactions on Mobile Computing* 15.6 (2016), pp. 1443–1456.
- [101] T. Barnett et al. “Cisco visual networking index (vni) complete forecast update, 2017–2022”. In: (2018).
- [102] R. Madlener, S. Sheykhha, and W. Briglauer. “The electricity-and CO2-saving potentials offered by regulation of European video-streaming services”. In: *Energy Policy* 161 (2022), p. 112716.
- [103] D. Kimovski et al. “Cloud, Fog, or Edge: Where to Compute?” In: *IEEE Internet Computing* 25.4 (2021), pp. 30–36.
- [104] B. Ramprasad et al. “Sustainable computing on the edge: A system dynamics perspective”. In: *Proc. of the 22nd Int. Workshop on Mobile Computing Systems and Applications.* 2021, pp. 64–70.
- [105] M. Yan et al. “Modeling the Total Energy Consumption of Mobile Network Services and Applications”. In: *Energies* 12.1 (2019).
- [106] S. Hu et al. “Modeling and Analysis of Energy Harvesting and Smart Grid-Powered Wireless Communication Networks: A Contemporary Survey”. In: *IEEE Trans. on Green Communications and Networking* 4.2 (2020), pp. 461–496.

- [107] Y. Xiao, Y. Song, and J. Liu. “Towards Energy Efficient Resource Allocation: When Green Mobile Edge Computing Meets Multi-Agent Deep Reinforcement Learning”. In: *2022 IEEE International Conference on Communications (ICC)*. 2022, pp. 4056–4061.
- [108] A. Mehrabi, M. Siekkinen, and A. Ylä-Jääski. “Energy-Aware QoE and Backhaul Traffic Optimization in Green Edge Adaptive Mobile Video Streaming”. In: *IEEE Transactions on Green Communications and Networking* 3.3 (2019), pp. 828–839.
- [109] S. Ling, Y. Baveye, and P. L. Callet. “Rate-distortion video coding and uncertainties: to be blindly chasing marginal improvement or to be greener”. In: *Applications of Digital Image Processing XLIII*. Ed. by A. G. Tescher and T. Ebrahimi. Vol. 11510. International Society for Optics and Photonics. SPIE, 2020, 115100F.
- [110] C. Wang and M. Zink. “Sustainable Cloud Encoding for Adaptive Bitrate Streaming over CDNs”. In: *2019 IEEE International Symposium on Local and Metropolitan Area Networks (LANMAN)*. 2019, pp. 1–6.
- [111] P. Borylo et al. “Energy-aware fog and cloud interplay supported by wide area software defined networking”. In: *2016 IEEE ICC*. 2016, pp. 1–7.
- [112] T. Tabata and T. Y. Wang. “Life Cycle Assessment of CO2 Emissions of Online Music and Videos Streaming in Japan”. In: *Applied Sciences* 11.9 (2021).
- [113] De Bosschere, Koen and Duranton, Marc. “Everything as a service”. eng. In: *HiPEAC vision 2021 : high performance embedded architecture and compilation*. HiPEAC, 2021, 206–211.
- [114] P. Suski, J. Pohl, and V. Frick. “All you can stream: Investigating the role of user behavior for greenhouse gas intensity of video streaming”. In: *Proc. of the 7th Int. Conf. on ICT for Sustainability*. 2020, pp. 128–138.
- [115] C. Díaz et al. “Energy- and Quality-Aware Video Request Policy for Wireless Adaptive Streaming Clients”. In: *IEEE Transactions on Consumer Electronics* 66.4 (2020), pp. 366–375.
- [116] M. A. Hoque et al. “Mobile multimedia streaming techniques: QoE and energy saving perspective”. In: *Pervasive and Mobile Comp.* 16 (2015), pp. 96–114.
- [117] A. Batmunkh. “Carbon Footprint of The Most Popular Social Media Platforms”. In: *Sustainability* 14.4 (2022).
- [118] T. Hossfeld. “What is the trade-off between CO2 emission and video-conferencing QoE?” In: *ACM SIGMM Records* (2022).

- [119] M. Fiedler, A. Popescu, and Y. Yao. “QoE-Aware Sustainable Throughput for Energy-Efficient Video Streaming”. In: *2016 IEEE International Conferences on Big Data and Cloud Computing (BDCloud), Social Computing and Networking (SocialCom), Sustainable Computing and Communications (SustainCom)*. 2016, pp. 493–500.
- [120] C. Herglotz et al. “Modeling of Energy Consumption and Streaming Video QoE using a Crowdsourcing Dataset”. In: *2022 14th International Conference on Quality of Multimedia Experience (QoMEX)*. 2022, pp. 1–6.
- [121] M. Mu et al. “A Scalable User Fairness Model for Adaptive Video Streaming Over SDN-Assisted Future Networks”. In: *IEEE Journal on Selected Areas in Communications* 34.8 (2016), pp. 2168–2184.
- [122] A.-N. Moldovan, I. Ghergulescu, and C. H. Muntean. “A novel methodology for mapping objective video quality metrics to the subjective MOS scale”. In: *2014 IEEE International Symposium on BMSB* (2014), pp. 1–7.
- [123] W. Sun et al. “Dynamic Backlight Scaling Considering Ambient Luminance for Mobile Videos on LCD Displays”. In: *IEEE Transactions on Mobile Computing* 21.1 (2022), pp. 110–124.
- [124] Y. Rubner, C. Tomasi, and L. J. Guibas. “The Earth Mover’s Distance as a Metric for Image Retrieval”. In: *International Journal of Computer Vision* 40.2 (2000), pp. 99–121.
- [125] J. Pang et al. “A new 5G radio evolution towards 5G-Advanced”. In: *Science China Information Sciences* 65.9 (2022), p. 191301.
- [126] Y. Ding et al. “Carbon emissions and mitigation potentials of 5G base station in China”. In: *Resources, Conservation and Recycling* 182 (2022), p. 106339.
- [127] D. Xu et al. “Understanding Operational 5G: A First Measurement Study on Its Coverage, Performance and Energy Consumption”. In: *Proc. of the Annual Conf. of the ACM Special Interest Group on Data Communication on the Applications, Technologies, Architectures, and Protocols for Computer Communication*. SIGCOMM ’20. Virtual Event, USA: ACM, 2020, pp. 479–494.
- [128] M. T. Sultan and H. El Sayed. “QoE-Aware Analysis and Management of Multimedia Services in 5G and Beyond Heterogeneous Networks”. In: *IEEE Access* 11 (2023), pp. 77679–77688.
- [129] A. A. Barakabitze, M. Liyanage, and A. Hines. “QoESoft: QoE Management Architecture for Softwarized 5G Networks”. In: *2020 IEEE International Conference on Communications Workshops (ICC Workshops)*. 2020, pp. 1–6.

- [130] B. Agarwal et al. “QoE-Driven Optimization in 5G O-RAN-Enabled HetNets for Enhanced Video Service Quality”. In: *IEEE Communications Magazine* 61.1 (2023), pp. 56–62.
- [131] ETSI. “5G; NR; Physical layer procedures for data (3GPP TS 38.214 version 16.2.0 Release 16)”. In: (2020).
- [132] E. F. Pupo et al. “Artificial Intelligence Aided Low Complexity RRM Algorithms for 5G-MBS”. In: *IEEE Transactions on Broadcasting* (2023).
- [133] C. C. González et al. “Dynamic Radio Access Selection and Slice Allocation for Differentiated Traffic Management on Future Mobile Networks”. In: *IEEE Transactions on Network and Service Management* 19.3 (2022), pp. 1965–1981.
- [134] C. C. González et al. “Network Selection over 5G-Advanced Heterogeneous Networks Based on Federated Learning and Cooperative Game Theory”. In: *IEEE Transactions on Vehicular Technology* (2024), pp. 1–16.
- [135] E. F. Pupo et al. “5G Link-Level Simulator for Multicast/Broadcast Services”. In: *2023 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*. IEEE. 2023, pp. 1–6.
- [136] N. Chukhno et al. “Efficient management of multicast traffic in directional mmWave networks”. In: *IEEE Transactions on Broadcasting* 67.3 (2021), pp. 593–605.

Appendices

Appendix A

Acronyms

4G	Fourth Generation Technology
5G	Fifth Generation Technology
6G	Sixth Generation Technology
ACR	Absolute Category Rating
ADASYN	Adaptive Synthetic Algorithm
AI	Artificial Intelligence
AL	Ambient Luminance
ANOVA	Analysis of Variance
AP	Access Point
API	Application Programming Interface
AR	Augmented Reality
AU	Action Unit
AUs	Action Units
BL	Backlight Luminance
BS	Base Station
CL	Content Luminance
CNN	Convolutional Neural Network
CO₂	Carbon Dioxide

CQI	Channel Quality Indicator
COVID-19	Coronavirus Disease 2019
DQX	Deterministic QoE Model
Extreme Gradient Boosting	XGBoost
FACS	Facial Action Coding System
FEC	Forward Error Correction
FCN	Fully Convolutional Network
FHD	Full High Definition
GD	Gaze Direction
GHG	Greenhouse Gas
GMEC	Green Mobile Edge Computing
GPU	Graphics Processing Unit
GWP	Global Warming Potential
HD	High Definition
HetNet	Ultra-Dense Heterogeneous Network
HLS	HTTP Live Streaming
HSPA	High Speed Packet Access
HTTP	Hypertext Transfer Protocol
ICE	Interactive Connectivity Establishment
ICT	Information and Communication Technology
ITU-T	International Telecommunication Union - Telecommunication Standardization Sector
ITU-R	ITU Radiocommunication Sector
ICKA	Improved Centered Kernel Alignment
IP	Internet Protocol
k-NN	k-nearest neighbors

KQI	Key Quality Indicator
LCD	Liquid Crystal Display
LLD	Low-Level Descriptor
LLS	Link-Level Simulator
LP	Laptop
LTE	Long Term Evolution
MCS	Modulation Coding Schemes
MFCC	Mel Frequency Cepstral Coefficient
ML	Machine Learning
MLP	Multi-Layer Perceptron
MLR	Multiple Linear Regression
MPAI	Moving Picture, Audio and Data Coding by Artificial Intelligence
MOS	Mean Opinion Score
NAT	Network Address Translation
NATS	Network Address Translation Traversal
NetEm	Network Emulator
NRS	Noise Reduced Speech
NR	New Radio
NSS	Non-Silent Speech
OS	Original Speech
P2P	Peer-to-Peer
PCA	Principal Component Analysis
PCC	Pearson Correlation Coefficient
PLR	Packet Loss Rate
QUIC	Quick UDP Internet Connections
QoE	Quality of Experience

QoS	Quality of Service
RAM	Random Access Memory
RB	Resource Block
RED	Redundant audio data
RES	Renewable Energy Source
RMSE	Root Mean Square Error
RTCP	Real time Transport Control Protocol
RTMP	Real-Time Messaging Protocol
RTP	Real-time Transport Protocol
SAM	Self Assessment Manikin
SDN	Software-defined networking
SIP	Session Initiation Protocol
SP	Smartphone
STUN	Session Traversal Utilities for NAT
SVM	Support Vector Machine
TCP	Transmission Control Protocol
TURN	Traversal Using Relay around NAT
TTI	Transmission Time Interval
TV	Television
UDP	User Datagram Protocol
VoIP	Voice over Internet Protocol
VR	Virtual Reality
WebRTC	Web Real-Time Communication
XGBoost	eXtreme Gradient Boosting

Bio

Gulnaziye Bingol earned a B.Sc. in Electrical and Electronic Engineering from Omer Halisdemir University in 2019. During her undergraduate studies, she broadened her academic perspective as a visiting student at the University of Oradea. Continuing her pursuit of advanced knowledge, she earned her Master's Degree in Internet Engineering from the University of Cagliari in 2021, where she distinguished herself by achieving the top score of 110/110. In 2022, Gulnaziye became a Ph.D. candidate in the Department of Electrical and Electronic Engineering at the University of Cagliari. Her research interests are primarily focused on improving Quality of Experience (QoE) in multimedia communications, exploring sustainability in digital technologies, and advancing the functionalities of Digital Twin and video communications. Her work aims to integrate technical proficiency with user-centric design principles to enhance digital communication platforms.

Acknowledges

This work has been partially supported by the PON “Ricerca e Innovazione” 2014-2020 (PON R&I) “Azione IV.4 Dottorati e contratti di ricerca su tematiche dell’innovazione”, assigned with D.M. 1062 on 10.08.2021.



UNIONE EUROPEA
Fondo Sociale Europeo



Ministero dell'Università
e della Ricerca



PON
RICERCA
E INNOVAZIONE
2014 - 2020



UNIVERSITÀ DEGLI STUDI
DI CAGLIARI

La borsa di dottorato è stata cofinanziata con risorse del
Programma Operativo Nazionale Ricerca e Innovazione 2014-2020, risorse FSE REACT-EU
Azione IV.4 “Dottorati e contratti di ricerca su tematiche dell’innovazione”
e Azione IV.5 “Dottorati su tematiche Green”