Analyse et modélisation de la qualité perçue des applications de visiophonie

Inès Saidi

Thèse

UNIVERSITE BRETAGNE LOIRE

THESE INSA Rennes
sous le sceau de l'Université Bretagne Loire
pour obtenir le titre de
DOCTEUR DE L'INSA RENNES
Spécialité : Signal, Image, Vision

présentée par

Inès Saidi

ECOLE DOCTORALE : *MATHSTIC*
LABORATOIRE : *IETR*

Analyse et modélisation de la qualité perçue des applications de visiophonie

Thèse soutenue le 28.02.2018
devant le jury composé de :

William PUECH
Professeur à l'Université de Montpellier / Président
Frédéric DUFAUX
Directeur de Recherche CNRS à Centrale Supélec Paris / Rapporteur
Alexander RAAKE
Professeur à l'Université d'Ilmenau, Allemagne / Rapporteur
Chaker LARABI
Maître de Conférences à l'Université de Poitiers / Examinateur
Lu ZHANG
Maître de Conférences à l'INSA de Rennes / Co-encadrante
Vincent BARRIAC
Ingénieur de R&D chez ORANGE / Co-encadrant
Olivier DÉFORGES
Professeur à l'INSA de Rennes / Directeur de thèse

Analyse et modélisation de la qualité perçue des applications de visiophonie

Inès Saidi

En partenariat avec

Acknowledgments

I would like to express my special appreciation and gratitude to my supervisors Vincent Barriac, Lu Zhang and Prof. Olivier Déforges for the continuous support of my Ph.D study. I would like to thank you for motivating me, encouraging my research and for allowing me to grow as a research engineer. Your advices on both research as well as on my career have been invaluable.

I would also like to thank my committee members, Prof. Alexander Raake from the Illmenau university and Prof. Frédéric Dufaux research director in CNRS Centrale Supélec Paris, for reporting my Ph.D thesis. I would like to thank Prof. William Puech from Montpellier University and Chaker Larabi from Poitiers University for examining my thesis. I want to thank you all for letting my defense be an enjoyable moment, and for your brilliant comments and suggestions, thanks to you.

Similar, profound gratitude goes to many people who helped me during this work: Pierre Henry for giving me the opportunity to join your team MOVE, for your trust and specially for your guidance and methodical advices to achieve the thesis objectives. Special thank to Laetitia Gros for sharing her expertise so willingly. My thanks go also to all the Orange Labs Technocenter Lannion team that helped me in recruiting subjects for my tests, collecting data for my Ph.D. thesis and putting at my disposal the subjective quality test laboratories.

A special thanks to my family. Words can not express how grateful I am to my mother and father for all of the sacrifices that you've made on my behalf. Your prayer for me was what sustained me thus far. I would also like to thank to my sister, brother and cousins for supporting me for everything, and especially I can't thank you enough for encouraging me throughout this experience.

Finally I thank my friends for providing support and friendship.

Abstract:

In a highly competitive environment, one of the key challenges for operators and providers of video telephony services is to ensure the highest quality of experience (QoE). There is a strong need for a measure that reflects users satisfaction and perception of these services. The audio-visual quality of a video call must be controlled to meet two main needs. The first concerns the planning of new technologies under development. The second is focused on the control of existing communications by assessing the quality of the services offered and evaluating them.

Two approaches are used to evaluate audio-visual quality: subjective tests by collecting scores given by participants on quality scales, after viewing and listening to audiovisual sequences and objective metrics based on automatic audio / video or audiovisual quality evaluation algorithms. Concerning telephony services, decades of research, standardization work and network exploitation, have allowed operators to master the automatic monitoring tools and to determine the representative metrics of voice quality. However, the metrics for measuring the audiovisual quality of a conversational services are not yet mature and not exploited by telecommunication operators.

The present work focuses on finding representative metrics of the perception of the video telephony and videoconferencing services quality. These objective metrics are calculated from the audio and video signals. Subjective tests are conducted to collect the judgment of service users on the perceived quality according to different levels of degradation. We studied the impact of network conditions (packet loss, jitter and desynchronization) on the QoE of a video call. The general principle is then to establish a correlation between the selected objective metrics and the perceived quality as expressed by the users. The results showed that new metrics of overall audiovisual quality that take into account the temporal aspect of video are more powerful than image quality based metrics. On the other hand, the use of a machine learning approach represents a solution to generate a global quality prediction model from the degradation metrics (blur, pixelization, image freezing, etc.).

Keywords: Quality of experience, audiovisual quality, conversational service, evaluation, subjective measures, objective metrics

Résumé:

Dans un contexte fortement concurrentiel, l'un des principaux enjeux pour les opérateurs et les fournisseurs de services de visiophonie est de garantir aux utilisateurs une qualité d'expérience (QoE) optimale. Il existe un fort besoin d'une mesure qui reflète la satisfaction et la perception des utilisateurs de ces services. La qualité audiovisuelle d'un appel vidéo doit être contrôlée pour répondre à deux besoins principaux. Le premier concerne la planification de nouvelles technologies en cours de développement. Le second est axé sur le contrôle des communications existantes en évaluant la qualité des services offerts.

Aujourd'hui, deux approches sont utilisées pour évaluer la qualité audiovisuelle : les tests subjectifs en collectant des notes données par des participants sur des échelles de qualité, après visualisation et écoute de séquences audiovisuelles et les métriques objectives basées sur des algorithmes automatiques d'évaluation de la qualité d'un signal audio, vidéo ou audiovisuel. Concernant les services de téléphonie, des décennies de recherche, de standardisation et d'exploitation des réseaux ont permis aux opérateurs de maitriser les outils de diagnostic et de déterminer les métriques représentatives de la qualité vocale. Cependant, les méthodes de mesure de la qualité audiovisuelle des services conversationnels ne sont pas encore matures et peu exploitées par les opérateurs de télécommunication.

Le présent travail est centré sur la recherche de métriques représentatives de la perception de la qualité des flux associés aux services de visiophonie et de visioconférence. Ces métriques objectives sont calculées à partir du signal audio et vidéo. Des tests subjectifs sont menés afin de collecter le jugement des utilisateurs du service sur la qualité perçue en fonction de différents niveaux de dégradations. Nous avons étudié l'impact des conditions réseau (perte de paquet, jigue et désynchronisation) sur la QoE d'un appel vidéo. Le principe général est ensuite d'établir une corrélation forte entre les métriques objectives sélectionnées et la qualité perçue telle qu'elle est exprimée par les utilisateurs. Les résultats ont montré que les nouvelles métriques de qualité globale audiovisuelle qui prennent en compte l'aspect temporel de la vidéo sont plus performantes que les métriques basées qualité d'images. D'autre part l'utilisation d'une approche machine learning représente une solution pour générer un modèle de prédiction de la qualité globale à partir des métriques de dégradation (flou, pixellisation, gel d'images, ...)

Mots clès: Qualité d'expérience, qualité audiovisuelle, service conversationnel, évaluation, mesures subjectives, mesures objectives.

Contents

List of Figures

List of Tables

Introduction

The first telecommunication permitting the transmission of speech between two people in real time was possible with the invention of the telephone in 1873 by Alexander Graham Bell. However, in a complete-distance communication, the need to add our own image in video quickly imposed itself. It was in 1972 that the CNET (French national center for telecommunication studies) established a first videophone link over broadband links between Paris and Lannion. The first consumer application was launched in 1984, during the "optic Fiber" experiment in Biarritz. Since then, with the deployment of new technologies of mobile networks, with increasing available bandwidth, the evolution of internet protocols and development of the devices (smart phones, cameras, PC, ...) video conversational services are becoming increasingly popular.

Video telephony, a technology that allows to see and interact with the interlocutor, offers different possibilities. "Point-to-point" is the closest thing to a phone conversation: two users are connected via video. "Multipoint" allows two or more people to take part in a video conference from a meeting room, a computer (in the office or at home), a smart phone or a tablet. The third option is broadcasting, which is a one-way signal transmission technique to a large number of customers. Broadcasting gives others the ability to access a meeting using software rather than hardware.

The challenge for operators and service providers is to offer to their customers the best possible Quality of Experience (QoE). The study of the QoE has been the subject of much scientific research to define it, to identify the impact factors and to investigate the methods to evaluate it. Monitoring the quality guides the actions of diagnosis and identification of the artifact causes. Thus, there is a strong need for automatic tools and metrics to evaluate the audiovisual quality of a video call as perceived by the end user.

The purpose of the work presented in this document is to contribute to studying the audiovisual quality perception in the context of a video call. We will focus on the essential impairments that may impact the user experience of a video telephony service which are related to network conditions. Subjective studies are conducted and objective models are evaluated in order to propose a toolbox for monitoring and diagnostic of the global quality of a visiophony service.

In **Chapter I** we will start with presenting the general context of the thesis by explaining the issues and the motivations of our research studies. We will define the principle technologies and network architecture allowing the development of a visiophony service. Finally, we will discuss the constraints we encountered to conduct

modulations on a consistent set of databases related to information collected from a service in use. Then, we will introduce the work methodology that we adopted throughout the thesis.

In **Chapter II**, we will present the state of the art and the researche conducted in the domain of the evaluation of audiovisual quality. First of all, it is essential to define the concept of QoE. Next, we will highlight the different impact factors that may influence the perception of the quality of a conversational service. Then, we will detail the two types of approaches existing to evaluate the quality: subjective and objective methods.

Chapter III, is dedicated to the presentation of the subjective experiments we conducted. We will present the methodologies, and the processes of the subjective tests we implemented. Then, we will analyze the results. We are interested in assessing the perception of video call service users under different conditions, and to constitute a sequences database to evaluate the performance of the objective quality metrics. We investigated the video, audio and audiovisual quality and asynchrony perception under two different situations: a non-interactive and an interactive conversational one. We analyzed the effects of network impairments (packet loss, delay) on perceived audiovisual, audio and video quality. We also evaluate the impact of experimental context and scene complexity on the quality perception in case of video calls. Furthermore, we propose new acceptability thresholds of audio-video asynchrony in video telephony context and study the effect of synchronization in the presence and absence of network degradation.

In **Chapter IV**, we will investigate in more details the perception of the audio/video synchronization in a specified study. Thus, we will show the results of two subjective tests conducted in order to better understand the influence of the time offset between the audio and the video media streams of video telephony contents in the presence of other impairments. We also compare between the subjective perception of quality and asynchrony in laboratory and in crowdsourcing contexts.

Once we collected different databases (from our subjective experiments and other public databases) composed of sequences with their subjective scores, we are able to apply objective metrics and conduct statistical and correlation studies. Thus, **Chapter V** is devoted to evaluating the prediction accuracy of the existent objective video, audio and audiovisual quality models. The main contribution of this chapter is to propose a representative global video quality metric that correlates best with the subjective perception. Furthermore, we will interest to no-reference single artifact based metrics by evaluating their performance in detecting different impairments that can occur for instance in a video conference call. We will associate each detected artifact to a specific cause or source (codec, network, rate adaptation ...) and will propose annoyance thresholds. Concerning the audiovisual quality we will consider the ITU-T G.1070 parametric computational model for point-to-point

videophone applications over IP networks.

The evaluation of the objective metrics allows us to determine the most accurate and representative of an audiovisual perception. Thus in **chapter VI** we will give methodology and primary results of applying machine learning algorithms on no-reference single artifact detection metrics in order to generate a global quality prediction model.

Finally, **chapter VII** concludes this thesis, and presents the different perspectives and directions for future research.

Chapter 1

Context

Contents

1.1 Thesis Issue

Over the years, multimedia applications have conquered many segments of the telecommunications industry. We are dealing today with multimedia services in many areas, starting with the various digital television systems, video-telephony, video-on-demand (VOD), Internet Protocol television (IPTV) or simply video-sharing services like YouTube or Dailymotion. Multimedia services represent an important part of the global IP traffic that is constantly growing. In the last statistics reported in [3], mobile video services will generate three quarters of mobile data traffic by 2020. Among the most popular multimedia services, the video conversational applications are in full development. In a competitive market, various Over The Top (OTT) players are emerging: Skype, Messenger, Facetime, WeChat, Duo, etc. For example, the statistics show that Skype has more than 300 million monthly active users [4] with 3 billion minutes per day spend on Skype video calls [5].

Fourth-generation mobile access networks (4G or LTE: Long Term Evolution [6]) defined by the 3GPP (3rd Generation Partnership Project) [7] allowed an increase in communication bitrate and bandwidth. As mobile operators already have significant experience in communication services, it is natural to take advantage of these developments in the access networks. Thus, mobile operators focus on video communications to leverage the video demand opportunity. Now, the launch of mobile voice over IP services is more and more via the integration of video and communication. This is referred to as ViLTE (Videotelephony over LTE) [8].

The development of these services and the end-to-end optimization of these systems are closely linked to the perception of quality by the user and his satisfaction

with the service rendered. In this sense, there is a strong need for a measure of user satisfaction and perception. Indeed, media service providers are increasingly interested in evaluating the performance of their services as perceived by end-users, in order to improve and better understand the needs of their customers. Network operators are also interested in this measure to optimize network resources and possibly (re)configure network settings to increase user satisfaction. Audiovisual quality measurement techniques are used to address two main cases. The first one concerns the planning of new telecommunication technologies under development, such as speech/video coding or speech/video denoising algorithms. The second one is focused on monitoring existing telecommunications by assessing the quality of the offered services and evaluating them.

There are several ways to get information about perceived quality. On the one hand, subjective evaluations are carried out in well equipped laboratories to investigate the perception of the end user. On the other hand, objective measures of quality are often used to study the measurable parameters of the whole system, describing the Quality of Service (QoS) in a technical way. However, these parameters cannot describe all the variables that influence the perception of quality on the end-user side. For this reason, Quality of Experience (QoE) was defined to better reflect the quality perceived by end users.

For telephony services, decades of research and standardization works (notably by ITU-T, IETF, ETSI, etc.) and the operation of networks have allowed to determine representative metrics of the quality perceived by the end-user (delay, audio quality, echo, noise, loss of information, etc.)[9] and to develop automatic tools allowing to know the performance of the network and its impact on end-to-end quality (such as passive probes to capture and analyze data flows in networks, or automatic systems used in mobile networks to perform tests that reproduce the experience of a client).

However, telecommunication operators and vendors have not strong expertise when it comes to ensure the supervision of these new videophone services. Indeed, there is a lack of experience to determine the right representative metrics and the associated thresholds to judge the acceptability of the quality of a service and to use tools with reliability and efficiency. The added value of the services offered by these operators lies largely in the fact that they are quality guaranteed. For example, 4G mobile access networks guarantee privileged processing of data transmitted on bearers marked by a Quality Call Indicators (QCI). QCI is a parameter present in the signaling at the establishment of an IP flow and making it possible to fix its main features, including its order of priority [10]. Thus, the QCI is equal to 1 for voice (low latency and packet loss with guaranteed bit rate), 2 for video (lower priority and packet loss but higher latency with guaranteed bitrate) and 5 for IMS signaling (absolute priority, without guaranteed bitrate), as opposed to the so-called Over The Top or OTT services, which use a non-prioritized best effort IP data transmission

channel. However, without any means of controlling (and ultimately proving) this end-to-end quality gain, this competitive advantage is partly ineffective.

This is why the development of appropriate methods for measuring and monitoring the perceived quality of these new services is becoming a major challenge for telecommunication operators. Beyond the complexity of access to data (separation of signaling and real-time data transport flows, security by data encryption, privacy), there is the question of the relevance of network indicators to represent the perceived impairments by the final user.

As mentioned above, the state of the art is rich in terms of voice or telephony quality measure. The main dimensions of perceived quality which are also found in regulatory systems, such as those applied in France for fixed telephony [11] and mobile telephony [12] are then:

- access to the service (service availability, call setup time),
- the intrinsic quality of conversational speech signal (generally characterized by scores between 1 for "very bad" and 5 for "excellent" called Mean Opinion Scores or MOS),
- the maintenance of the call (efficiency of cell changes in mobile networks, hung up prematurely).

Voice quality metrics are well mastered, most often standardized (notably by the ITU, in the E, G and P series of recommendations [13]), and the methods for evaluating them are proven by long years of experience. The most emblematic method, known as the Perceptual Objective Listening Quality Assessment (POLQA) [14, 15] , concerns the measurement of mean opinion scores from an analysis of the audio signal received from a transmission chain and its comparison with the corresponding reference in the sending side.

These different metrics are integrated into test or supervision tools, manufactured by some specialized companies (for example the French companies Witbe, IP-label and Opale Systems, but also, among others, Rohde & Schwarz, Opticom, Keysight, Viavi or Exfo) and sold (often very expensive) to telecommunication operators.

The voice quality measurement tools, thanks to a long experience, are now embedded reliable metrics useful to diagnose and correct problems. The situation is absolutely different for conversational audiovisual services, for several reasons.

- Technical complexity of measuring video quality. The video content is far more complex than the speech because of the amount of spatial and temporal information it contains.

- Influence of the coding and the transmission in IP networks.
- Multiplicity and complexity of used terminals and screens of different size (PC, smartphone, TV, etc.)

However, the operational needs are beginning to emerge. For many years, test tool manufacturers have been offering solutions dedicated to the supervision of audiovisual streaming services and trying to adapt them to the problem of conversational services. The technical difficulties mentioned above indicate that, most of the time, these tools are specialized on a service available with a given image format and on a given terminal model. In addition, the absence of universally recognized or standardized metric results in an abundance of proprietary methods that are incomparable among themselves and whose correlation with the perception of the end user is questionable.

Operators are therefore reluctant to embark on major investments whose reliability is not proven. There are certain standards, but telecommunications operators urgently need tools that are adapted to their needs. This will be made possible if they have the most possible generic knowledge about representative metrics of the quality perceived by their customers and how to estimate them automatically, but also if this knowledge is shared with the ecosystem, especially manufacturers of measurement tools. The latter have every interest in being able to justify the relevance of their technical approach on the basis of results published in scientific journals or in standardization bodies. The final benefit of this work goes to the user of the services, to which we can provide a verifiable quality of service.

Based on the rapid growth of the use of these services, as well as on the available state of the art from telephony and in the field of audiovisual broadcasting services, the elaboration of reliable, long-lasting and recognized solutions for monitoring conversational services is necessary.

1.2 Technical context

1.2.1 ViLTE

Long-Term Evolution (LTE) is a standard for high-speed wireless communication for mobile devices and data terminals, based on the 2^{nd} and 3^{rd} mobile network generation technologies. LTE networks can deliver mobile broadband with greater data capacity and lower latency. However, as there is no circuit-switched voice domain in LTE, the mobile industry has adopted a globally interoperable IP-based voice and video calling solution for LTE, known as VoLTE, which also enables development of new innovative communication services. VoLTE is a foundation for a modern user experience including services like HD voice, video calling, HD conferencing, IP messaging and contact management, as well as new innovative services.

Based on the IP Multimedia Subsystem (IMS) core network, voice services over LTE can be enhanced to a high quality conversational video calls by adding a video capability, providing users with synchronized full-duplex voice and videostreams. With the video communication over cellular LTE network (ViLTE), users can theoretically make one-to-one or one-to-many video calls, switch to video at any point during a call, and drop video at any point to continue with just voice.

ViLTE represents an opportunity for operators to offer a high-quality voice, video, and rich multimedia experience to end users, in order to compete against OTT applications. In addition to the quality of experience, ViLTE is supposed to allow operators to provide security and flexibility, what the OTT video apps cannot guarantee. However, mobile operators' deployment of ViLTE applications has not been widespread with only 16 launches as of August 2017 (in comparison, there are 113 VoLTE launches and 621 LTE Launches). Therefore, it is important to identify the factors and challenges that block the adoption of ViLTE.

To fully exploit the potential of ViLTE, the services provided by different operators must be interconnected. In fact, subscribers must be able to reach others without having to worry about whether the called party is subscribed to the calling party's network. As ViLTE possesses a very diverse set of parameters, it is more challenging to interconnect ViLTE services than to interconnect VoLTE services, reducing the benefit of ViLTE for operators through additional cost and complexity. In addition, ViLTE is a video calling service with a guaranteed quality and consequently may impact stability of the network unless network resources are planned carefully. This requires consideration and time, which increases complexity of the service whose demand is not widespread. Lastly, it may be difficult to coordinate interconnect charging and troubleshooting in interconnection scenarios due to organizational reasons.

Country	Operator
Argentina	Movistar
Australia	Telstra
Brazil	TIM Brasil
Indonesia	Smartfren
Macau	CTM
Slovakia	4ka
Turkey	Turkcell
Argentina	Personal
Czech Rep	T-Mobile

Table 1.1: ViLTE deployment status (source: GSA)

1.2.2 RCS IP video call

Rich Communication Services (RCS) is a functionality of IMS defined by the GSM association and offers to the customer a set of innovative features to complete the basic functionality offered by SMS: the customer can initiate individual or group chat sessions and have rich voice and video calls. RCS combined with VoLTE can bring new opportunities for operators and enable them to compete against OTT players.

RCS has the advantage of inter-working between networks and devices, unlike OTT services (no application is required on the caller and callee sides). For operators with an IMS network, RCS compatible devices connect through appropriate access such as Wi-Fi, LTE and 3G. The device must then register and authenticate with the ability to use the RCS messaging service.

Once the device is registered, the IMS network routes all RCS messages to the RCS messaging service and to other IMS networks. RCS services include standalone messaging, 1-to-1 and group chat, file sharing, sending audio messages, enhanced voice communication before and during the call, geolocation. RCS offers better quality thanks to the possibility of integrating QoS and Resources Management.

The RCS IP video call service allows to a user under only 3G or LTE network cover to use the ViLTE service which guarantees a quality of service during the video telephony call on IP and the continuity of the service with failover on the circuit domain if the HSPA coverage or LTE is no longer available during the videophone session. If the usage is switched on the 2G radio the call continues with only the voice component. If the call is switched to the 3G radio, the video call can continue in circuit mode.

1.2.3 WebRTC

WebRTC is an open source standard for the web multimedia conferencing systems published by Google in 2011 [16]. It is a technology specified by the World Wide Web Consortium (W3C) and the Internet Engineering Task Force (IETF) to provide real-time communication capabilities to media-capable end points (e.g. browsers, native applications) [17]. It represents an HTML5 extension for real-time communications, enabling live media communications between two or more parties using standardized web technologies.

It has been developed to enable communication with only few lines of JavaScript code, without any plugins, and it is supported in browsers such as Chrome, Firefox and Opera [18]. IETF has defined a set of protocols to exchange data (voice, video, text, etc.) in peer-to-peer mode, including NAT traversal protocols with ICE (Interactive Connectivity Establishment). A key issue is that the signalling protocol between end points is not fully specified and left to service providers. The

only signalling constraint is to rely on JavaScript Session Establishment Protocol (JSEP) [19] which makes use of Session Description Protocol (SDP) to exchange media capabilities and other parameters (e.g. ICE candidates).

The standardization goal is to define a WebRTC API that enables a web application running on any device, through secure access to the input peripherals (such as webcams and microphones), to exchange real-time media and data with a remote party in a peer-to-peer fashion. Details on the WebRTC architecture and principle APIs are given in Appendix 7.

A comparison between the listed above video call technologies is summarized in Table 1.2.

Technology	WebRTC	RCS	ViLTE
Definition	Web Real-Time Communications	Rich Communications Services	Video over LTE
Developed by	Open Source, Web Developer Community	Telecom Standard, GSMA	Telecom Standard, GSMA, 3GPP
Standard	API: W3c, RTCWeb(transport) IETF	IR.84 v12.0, IR.74 v2.0	IR.94 v12.0, 3GPP (TS 26.114 v15.0.0)
Providers	OTT & operators	Network Operators	Network Operators
Prerequisites	None	IMS	IMS
Device	Platform and Device independent: Web, Mobile	Modern and compatible smartphones	Modern smartphones
Coverage	Internet Data	Cellular (3G, 4G)	Cellular (4G)
Audio codecs	Opus (RFC 7874), AMR, AMR-WB, G722	AMR	AMR, AMR-WB
Video codecs	VP8,VP9, H.264	H.263, H.264	H.264, H.265
RTCP	Sender Report (SR) & Receiver Report (RR)	SR & RR	SR & RR
Adaptation Error recovery	SAVPF(Secure Audio Video Profile Feedback)	AVP (Audio Video Profile)	Extended AVP Feedback

Table 1.2: Overall WebRTC - RCS - ViLTE comparison

1.3 Reflection

The aim of this thesis is to study and propose representative metrics of perceived quality associated with video calling and video conferencing services. These metrics are to be determined using information from the audio signal and video, but also by analyzing the service elements accessible at the terminal or network equipment (service platforms, in particular). Our first focus was to model the perceived quality of the audiovisual services using technical information collected at the terminals and networks using datamining methods.

Thus, the research studies have to rely on offline methods, to collect and analyze usage and perception data in large quantities from the users themselves. This requires access to the technical and usage data of a video conversational Orange service with a large number of customers. To achieve this goal we determined three possible tracks:

- ViLTE: possible deployment in addition to VoLTE in some countries.
- Web RTC: several internal projects giving place to experiments.
- Orange Libon: OTT solution already widely deployed, but not yet in video context.

A study conducted in 2014 by Orange Labs on digitalization and unified communications has shown that WebRTC solutions would strengthen Internet and data access services for Small and Medium size Entreprises (SME). Following this first study and technical experiments, it was decided, as part of a research project, to conduct an experiment with Orange Ivory Coast's teams on an application developed internally named "PLACE" to:

- technically test the solution in Ivory Coast,
- measure and validate the bit rates required for good audio and video quality,
- evaluate perceived quality for different communication scenarios,
- evaluate the quality of the service "PLACE".

We consider "PLACE" for our WebRTC option solution.

Libon [20] is a voice-over-IP communication application developed by Orange. It offers High Definition (HD) voice calling out to mobile numbers, voice mail and online chat messaging features on iPhone and Android clients.

Between ViLTE, WebRTC and Libon, our choice of the adequate solution for our study was based on several criteria. First, it is essential to precise if the service is existent, how many users use it and if it is commercialized. Then it comes to determine the availability of the technical information (from network and terminal),

the client perception information (test results, embedded agent, regular feedback via polls, ...) and the way to access to these information (confidentiality/right of use, access rights, organization of the database, ...). Table 1.3 summarizes the comparison between the services.

	ViLTE	WebRTC	Libon
Voice service	Yes	Yes	Yes
Type of service	Operator video call	collaborative experimentation	Commercial
Number of users	0	"Place": 40	Around 1 million
Video service	No	Yes	No
Technical information	Yes	Yes	Yes
Network information	IP Probes, SIP IP and RTP metrics	No	PF data collection
Terminal information	No	GetStats	Collection of usage data + random questionnaire
Type of collect	No	"Place" 1 ticket per call	Centralized collection server
Identification data	No	Yes	Yes
Application polls	No	No	Yes
Access to data	No	Yes (for "Place")	Partially
Confidentiality	No	possibility of anonymizing the data	very important

Table 1.3: Overall WebRTC - RCS - ViLTE comparison

After analyzing the possible solutions to apply QoE modeling of video telephony and visioconference services, we notice that there is not a satisfactory solution.

- For ViLTE: difficulties in its deployment in France; no client data available
- For WebRTC: limited number of users of Orange applications (Place)
- For Libon: no video service offered.

In order to increase the accuracy of a predictive model, it is essential to have a fairly consistent database. Taking as the first selection factor the size of the available database, we chose the Libon service. On the other hand, despite the fact that Libon is only providing a voice call service, it had a prototype of the video component that

was considered to be integrated later (this was at the beginning of the thesis). We have launched our statistical studies on voice call data with the objective of applying the same methodology with video calls once they are deployed. This study is a first step for using huge network data and demonstrating opportunities offered by big data and statistical tools. The next steps are to develop this kind of methodology with new data sets and to share models via existing Orange research tools. Detailed description of the collected Libon database and the obtained results are presented in Appendix II. Statistical and correlation studies show that after pre-treatment of the database and elimination of the outlier information, we are face to insufficient volumetry. On the other hand, the database is fragmented: difficulty to find users with close profile (same country, access network, codec, ...).

Given these impossibilities and technical constraints that have made the integration of visiophony services is still in development and that there is not, at the beginning of the thesis, a usable Orange service for video call data collection, we went back to classic solutions namely the conduct of subjective test companies. Thus, we studied the impact of network degradation on the video call service end user perception and we investigated the representative objective metrics for the global quality and for the detection of possible video artifacts with the determination of the corresponding annoyance thresholds.

In next Chapter, we will present the state of the art in audiovisual quality evaluation before presenting our research works and contributions in the followings Chapters.

Chapter 2

Evaluation of audiovisual quality: state of the art

Contents

Introduction

With the rapid development of broadband telecommunication technologies and the expansion of mobility (3G, LTE, 5G and WIFI), various applications (e.g. video telephony, video-sharing and e-learning) have been created to complement face-to-face conversations. They are usually low-cost, compatible with mobile devices, capable of transmitting multimedia contents, thus have achieved widespread popularity. However, the quality of service (QoS) of these new applications is usually not guaranteed. In practice, with IP-based networks, there is no guarantee that the streams transmit without errors. Many processes in the supply chain may degrade the perceptual quality. Meanwhile, telecommunication operators are competing to offer an optimal user experience to their customers. Their main goal is to establish a trade-off between the user satisfaction and the available network resources.

Thus, special attention is paid to assess the quality of experience through the development of tools and the implementation of evaluation methods. For audiovisual service providers, the Quality of Experience (QoE) is particularly studied through the perception of the quality of the media (ie the quality of the audio and/or video signals returned to the user). Perceived quality, and more broadly QoE, becomes a key element that must be studied and measured.

In this chapter, we will present the state of the art and the researches conducted in the domain of the evaluation of the audiovisual quality. First of all, it is essential to define the concept of Quality of Experience. Next, we will highlight the different impact factors that may influence the perception of the quality of a conversational service. Then, we will detail the two types of approaches existing to evaluate the quality: the subjective and the objective methods.

2.1 Perception of quality: definitions and concepts

In order to take full account of the impact of QoE in current and future conversational services, it should first be necessary to define this notion precisely. Because of the multiplicity of criteria that can be taken into account, it is difficult to define a concept as broad as the QoE.

The term QoE appears in many works. In [21] Kalevi Kilkki proposed a generic definition of the QoE: *the basic character or nature of direct personal participation or observation.* We can find an extensive study of the meaning of quality and experience in [22]. The writer has defined the **experience** as *the individual stream of perceptions (of feelings, sensory percepts and concepts) that occurs in a particular situation of reference.* Therefore, experiencing may have direct relation with feelings not only with pragmatic concepts. The **quality** is presented as *the judgment of the user based on those feelings and his expectations.* These definitions are in coherence with the latest definition of QoE presented by QUALINET through its white paper [23]:

QoE is the degree of delight or annoyance of the user of an application or service. It results from the fulfillment of his or her expectations with respect to the utility and/or enjoyment of the application or service in the light of the users personality and current state.

This definition is now considered as the universal one adopted by the experts of the domain and by the ITU-T Study Group 12. In this definition, the term "personality" is used to mean "the characteristics of a person who count for a coherent pattern of feelings, thoughts and actions" [24]. The term "current state" is used to mean "temporal or situational changes in a person's feelings, thoughts or behavior". It can be noted that the current state (relaxed, happy, stressed, etc.) is both an influential factor of the QoE, but also a consequence of the experience.

Finally, two words can draw our attention in the above definition of the quality of experience:

- Application: software and/or hardware that allows the interaction of a given content. This may include entertainment, information, documentaries...
- Service: Use that we can make of something.

In the context of conversational services, QoE can be influenced by many factors such as the type of service or the service itself, the content, the network, the broadcast material, the application used, the context of use, expectation, past experience, etc [25]. In the following section we will define and classify the different factors and parameters that impact the user perception of a conversational audiovisual service.

2.2 Influence factors

Existing studies have proposed classifications of factors impacting QoE, often termed QoE influence factors, for various types of multimedia services [26, 27, 28, 29]. While a factor is a characteristic which influences QoE, it is not a part of the perceived QoE itself. Extensive work on factor classification has been performed by S. Jumisko-Pyykkö [30] in the form of a User-Centered Quality of Experience (UC-QoE) model where characteristics of the user, system/service, and context of use are identified as contributing to different experiential dimensions of QoE.

In the context of communication services and applications, the factors influencing QoE are defined by QUALINET in its white paper [23] as any characteristic of a user, system, service, application, or context whose actual state or setting may have influence on the Quality of Experience for the user. Thus, there are many factors that have an impact on perceived audiovisual quality. These factors depend on the application, network technology, user terminal, etc. In [31] the authors classified the impact factors of general networked services and electronic communication services and applications into three categories:

- Human influence factors: any variant or invariant property or characteristic of a human user (demographic and socio-economic back-ground, physical and mental constitution, user's emotional state).
- System influence factors: properties and characteristics that determine the technically produced quality of an application or service.
- Context influence factors: are factors that embrace any situational property to describe the user's environment.

More specifically, in the context of videoconferencing services, the impact factors of QoE are studied in [32]. The authors consider that the QoE has three dimensions: System, User and Context. The System dimension is composed of the technical parameters which are the application QoS, the System QoS and the Network QoS. The

context one represents the socio-cultural, the situational and the interactional conditions. The user dimension defines the role played by the user in the communication (case of group conversation).

Considering all the studies conducted in defining and listing the possible factors impacting the QoE, we propose in the following section our own classification of these factors in video conversational services, depending on the source of the degradation.

2.2.1 Network conditions

Network conditions belong to the category of System influence factors. Network design and management are a key element in the quality of a video-conference call. Typically, network conditions include packet loss, delay, jitter, and bandwidth factors. The effect of these parameters on the perceived quality depends essentially on the type of the multimedia application. Several studies investigated the impact of network impairments on the QoE in different contexts [33, 34, 35]. In [36] authors showed that video content characteristics, the encoding scheme and the error concealment, affect the visibility of artifacts caused by network errors (packet loss and jitter).

For video-conference applications, which are real-time services, the packet loss rate and the bandwidth are the most important network parameters. Since there is interactivity, delay and jitter also play an important role, adding echo and loss of audio/video synchronization.

2.2.2 Applicative characteristics

The content type of the video and the audio streams has an obvious and strong impact on the overall perceived quality. For example, the luminance level, the spatial and temporal complexities of the scenes and the ambient noise of the room have a significant impact on quality, especially when there are other factors, such as very low bit rate encoding and/or packet loss in the network. Source parameters that depend on the characteristics of the sequence, such as the nature of the scene (eg, amount of movement, details, texture, color, contrast, frame size, noise level, audio frequency, etc.) also have an impact on Human perception of the quality of the video.

Encoding or compression parameters are important content factors. For stronger video compression, these will usually give visible blocking (rectangular shaped) distortions and blurring, whereas wavelet based techniques mostly give blurring distortions as in JPEG 2000. As examples of these parameters, we can mention the type of used codec (H.264, HEVC, MPEG-2, etc.), number of bits per sample, bit rate, frame rate, number of layers in the case of layer coding, etc [37, 38].

For audio, the coding also depends on the content type and service. Several lossy compression codecs are used for audio media. Among the applicative parameters influencing the audio quality we can mention the quality improvement techniques such as: echo cancellation, Silence detection and suppression, error correction and interleaving.

2.2.3 Context

The context characterizes the environment in which the user makes his video call. The context influence factors have been considered in different studies Here we consider as context factor the one who belong to the following categories:

- Psychological context and Sociocultural background which refer to the profile of the user and its emotional state. These factors are highly complex because of their subjectivity and strong relation to internal states and processes. This makes them rather immaterial and therefore much more difficult to comprehend. Several studies were interessted in investigating the importance of these human factors and their impact on the QoE [39, 40, 41, 42, 43]. In some empirical studies, subjective and physiological indicators are taken into account in QoE evaluation.

- Spatial and temporal context including user location (home, work, outdoor, indoor, airport ...), time of day or week (morning, late at night, weekend ...). The location and space, including movements and transitions between locations have an important impact on the quality of a video communication. In [44] authors give a detailed explanation of these factors.

- Use case (motivation): interview, call with friend, meeting. activity(walking, stable ...)

We can cite, as context parameters, the ambient noise level of the room, the loudspeakers / microphone used, the capabilities of the decoder/computer... This type of parameters is difficult to measure and the most often uncontrollable.

2.2.4 Impact of desynchonization

When transmitting data, it does not really matter when a packet is delayed from arriving. However, with real time conversational communication, the overall delay between the image and the sound is extremely important. The time that elapses between a person, who says something, and another who watches and listens to what has been said, should be as low as possible. Otherwise, a lip desynchronisation will be noticeable.

Large delay values result in loss of interactivity. In the case of real-time communications, from the application point of view, the delay is generated due to analog/digital conversion, signal compression and decompression, packet encapsulation,

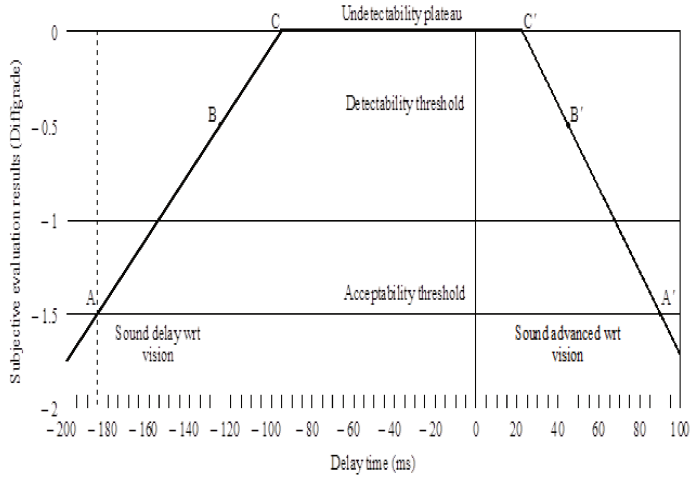


Figure 2.1: Detectability and acceptability thresholds for sound/image asynchronism, as per ITU-T BT.1359

the network interface, the propagation time and latency in the network, and the time required for the dejittering of the stream at the receiver side.

Managing the synchronization of audio and video signals is a central concern in the broadcasting of audiovisual content. Indeed, several studies have shown that a perceptible difference between the transmission times of the sound and image components of an AV signal is inconvenient for the user. It has been found that for a television context (TV news), desynchronization is perceptible from -45 ms (sound in advance) and +125 ms (late sound) and unacceptable from -90 ms and +185 ms [45, 46]. An illustration of the trays of perceptibility and acceptability is provided by Figure 3.9 below. On the other hand, perceived quality degrades rapidly when desynchronization increases. Specifically, desynchronization would be perceived as annoying from 150 ms in advance of sound on the image [45, 47].

The telecommunications world is more interested in the impact of the entire transmission chain on the original signal, ie the signal transmission (through the coding/decoding the transmission channel) to the rendering terminal. Figure 2.2 illustrates the path of the audiovisual signal from its production to the perceived final quality. The quality of the signal is then considered from the point of view of the user of audiovisual reproduction services. Thus in our thesis study we didn't consider the impact of the context factors, but we rather investigated the audiovisual quality perception under different network, applicative and synchronization conditions. Results and analysis will be discussed in Chapter 3.

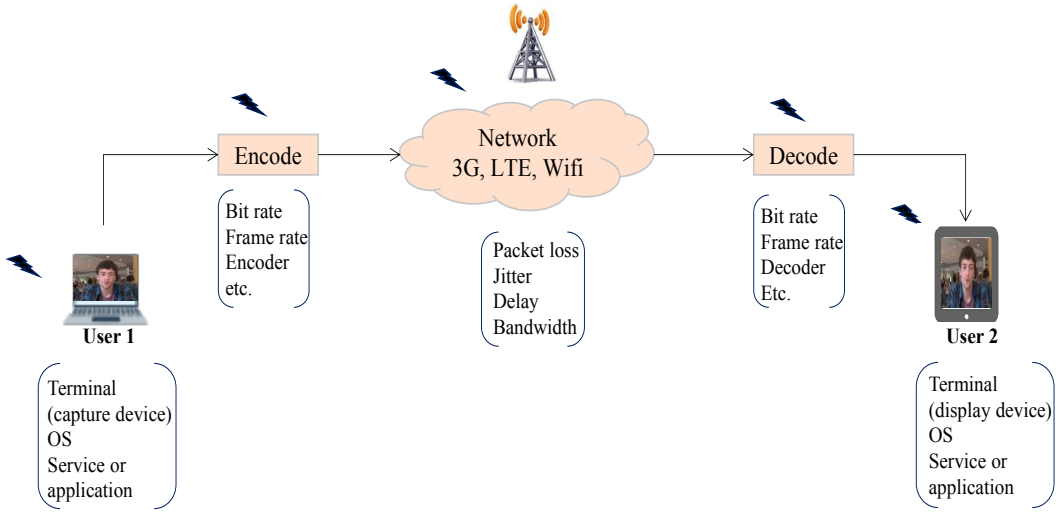


Figure 2.2: End-to-End communication chain of a video-conference service

2.3 Subjective evaluation methods

Subjective assessments is the most accurate way to measure the quality of a multimedia stream. In subjective experiments, a number of subjects (observers or participants) are invited to attend a set of tests and to judge the quality of the media or the inconvenience caused by the distortions. The average of the values obtained for each test sequence is known as Mean Opinion Score (MOS). In general, subjective assessments are costly and time-consuming. As a result, the number of experiments that can be carried out is limited and, therefore, an appropriate methodology must be used to make the best use of resources. In the following sections we will define the different protocols used in subjective audiovisual, video and audio tests.

2.3.1 Audiovisual quality

Although it is widely accepted that the perceived media quality is a multidimensional phenomenon, the vast majority of evaluation methodologies assume that the quality of an audio and/or video signal can be described by a scalar on a one-dimensional quality scale. The notion of quality is then reduced to a general impression or overall quality, integrating all the underlying dimensions. The scores collected for each individual are then averaged, for a given test sequence, on all participants. The average score of opinion or the MOS (Mean Opinion Score) obtained, will then determine the level of quality of the evaluated signal.

The ITU has made recommendations for subjective testing procedures. In general, these recommendations focus on evaluating a single modality, audio or video,

at a time. For example, ITU-T P.800, ITU-T P.805, ITU-T P.806, ITU-T P.835 [48, 49, 50, 51] are recommended for voice quality assessment, Recommendations ITU-R BS.1284-1 [52], ITU-R BS.1534-1 [53] and TU-R BS.1116-1 [54] allow the evaluation of audio quality while ITU-R BT.500-13 [55] and ITU-R BT.1788 [56] are dedicated to video quality assessment.

Some standards also suggest methods for evaluating a given modality (audio or video) in an audiovisual context: ITU-R BS.775-3 [57] and ITU-R BS.1286 [58] allow the evaluation of multichannel audio (digital television broadcasting) and audio systems, in general, in the presence of an accompanying image.

In our context of videoconferencing services, in general multimedia systems, the recommendation ITU-T P.910 [59] provides methods for evaluating video quality. Only two standards are dedicated to the subjective evaluation of audiovisual quality for an interactive (ITU-T P.920, [60]) or non-interactive (ITU-T P.911, [61]) context.

ITU-T P.911 proposes audiovisual quality (AV) assessment methods for non-interactive multimedia applications (passive context of listening and viewing: TV, multimedia, etc.). The quality judgment is made on a single scale at the end of the visualization and the listening of each audiovisual test sequence. Four methods are proposed under this standard; they are described in the following paragraphs.

Absolute Category Rating (ACR) method

The ACR method, also known as the Single Stimulus Method (SSM), consists of assigning a quality score after each visualized/heard AV sequence. The given score should reflect the participant's view of the perceived overall audiovisual quality, ie the combined audio and video quality. This evaluation is performed on a five- or nine-point (interval) categorical scale that is explained by five items (Excellent-Good-Fair-Bad-Poor). An illustration of the recommended scales is given in Figure 2.3. The ACR method is an inexpensive method from the point of view of its application, treatment and analysis of the results. It also has the advantage of being able to qualify test systems and obtain their ranking according to the level of quality associated with them.

Degradation Category Rating (DCR) method

The DCR method proposes a presentation of the AV test sequences in pairs. The sequences constituting the pair are identical to the difference that the first one is always presented without degradations (reference) while the second is processed by the system to be evaluated (and therefore liable to involve degradations). The processed sequence is always presented after the reference. Only the processed sequence is evaluated by the participants in comparison with the reference condition.

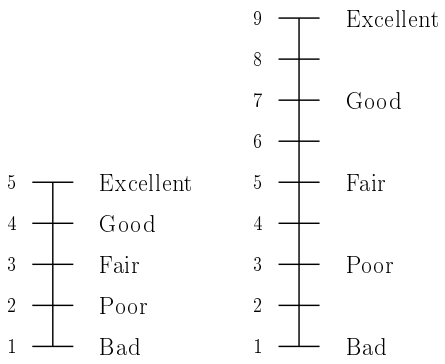


Figure 2.3: Scale of quality assessment (MOS) at 9 and 5 levels.

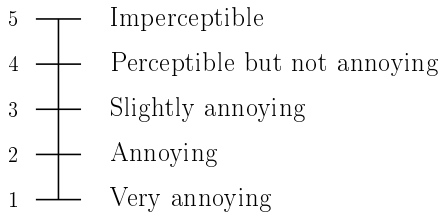


Figure 2.4: Scale of quality degradation (DMOS) at 5 levels.

The scale of assessment here corresponds to a scale of perception of the degradation (DMOS) as presented in Figure 2.4. The duration of the test sequences and the voting time are identical to those recommended in the ACR method. The main advantage of this method is that it allows a rapid qualification of the level of discomfort associated with certain degradations generated by the systems under consideration.

Paired Comparison (PC) method

The PC method consists in presenting two identical sequences, with the difference that each sequence is treated by a different test system. The reference sequence (without degradation) can also be included as an additional test system. All combinations of sequence pairs A, B, C, etc. should be evaluated (AB, BA, CA, etc.) and presented in the two possible orders (AB, BA, etc.). The overall AV quality judgment is here expressed through a judgment of preference for one or the other sequence of the pair. This judgment is made after the presentation of each pair. This method is particularly recommended for the comparison of quasi-equivalent and/or high-quality systems. The recommended duration for the test sequences is approximately ten seconds, the duration of the voting time must be less than or equal to ten seconds.

Single-stimulus continuous quality evaluation (SSCQE) method

A final method, the SSCQE method, is a continuous assessment method to collect the evaluation of the participants during the visualization of the test sequences for which the quality level fluctuates. The testers report their judgment by means of a slider which can be moved along a continuous scale. This allows a score to be assigned between 0 and 100 where 100 represents a perfect quality. The scale is divided into five equal segments corresponding to the five-point quality scale, the items characterizing the different levels are identical to those of the ACR method. No reference is given to serve as a basis for subjective evaluation. The duration of test sequences proposed is much greater than the previous methods. This can be between three and thirty minutes.

The choice of one method over another will be guided according to whether the fixed objective corresponds to a fine discrimination between several systems, to a qualification of systems or to a detection of degradations.

ITU-T P.920 [60] provides recommendations for the evaluation of audiovisual communication services (interactive multimedia applications such as videoconferencing). The proposed communication tasks ($<$ 5 min) should encourage participants to communicate in the most natural possible way and remain focused on the audiovisual media. ITU-T P.920 describes different communication scenarios to engage the participant in the activity: question/answer set, comparison of stories or images, etc. The evaluation of audiovisual quality is carried out on the basis of a multi-criteria approach. In particular, it is possible to ask the participants to judge the overall audiovisual quality but also the audio and video qualities judged separately. In this case the assessment scale is the ACR one with five-points level. It is possible to ask the participants to assess the effort needed to interrupt using the categories: No Effort, Minor Effort, Moderate Effort, Considerable Effort, or Extreme Effort. The communications difficulty and acceptability of communication can be assessed using a binary choice: Yes or No.

2.3.2 Video quality

For subjectively evaluating the video quality of multimedia applications, the test protocole is described in Recommendation ITU-T. P.910 [59]. This document provides information on video display conditions, selection criteria for observers and test equipment, evaluation procedures, and methods of data analysis. Before choosing the method to be used, we must take into account the application and the objectives of the evaluation.

According to the ITU-T. P.910, there are two categories of subjective assessments:

- Quality assessments: the scores given by the participants are on a quality scale, ie, the quality of the video displayed is good or bad. These evaluations are used to evaluate the performance of the systems used in optimal conditions.

- Depreciation tests: judgments made by subjects are on a scale of value, ie, the distortions of the displayed video are visible or imperceptible. These evaluations are used to assess the ability of systems to maintain video quality under non optimal conditions. These methods are often used to measure quality degradation caused by coding or transmission patterns.

The assessment scales, for quality assessment or for the evaluation of degradation, may be continuous or discrete. Judgments can also be categorical or non-categorical, adjectival or numerical. Depending on how presenting the video sequence, evaluation methods can be classified as a single or double stimulus. In the simple stimulus approach, only the test sequence is presented, while in the double stimulus method, a pair of sequences (test sequence and the corresponding reference sequence) are presented together. The evaluation procedures of ITU-T Rec. P.910 are as for P.911: ACR, DCR and PC.

2.3.3 Speech quality

ITU P.800 [48] describes the methods and procedures for conducting a subjective assessments of speech transmission quality. The most commonly used method is Absolute Category Rating (ACR). The Degradation Category Rating (DCR) is also used on some occasions. Subjective assessment is usually performed under an acoustically treated room.

ACR tests are most commonly used to assess the integral quality of speech (ITU-T Rec. P.800 [48]). In this type of test, a group of listeners evaluates a series of audio files (voice) using a five-value scale, without having to listen to the original sequence.

When good quality speech samples are evaluated, the ACR method tends to be insensitive to small quality degradations. The DCR degradation category assessment procedure, which relies in particular on a disturbance scale and a high quality reference, seems to be suitable for evaluating good quality speech. The subjects noted the level of degradation and discomfort by comparing with the original speech signal. In order to standardize subjective tests, ITU P.800 defines detailed conditions such as test material characteristics and the test environment. Subjective tests are normally performed in a controlled laboratory area, double-walled, soundproofed room.

2.4 Objective evaluation methods

The current methods of quality assessment are mainly standardized by the International Telecommunication Union (ITU)[62] allowing a comparison of results from different laboratories. The ITU-T G.1011 Recommendation provides a reference guide to QoE assessment methodologies [63]. According to the ITU studies [63, 64], objective metrics may be classified into five main categories depending on the type of input data:

- Media-layer models use the audio or video streams to evaluate the perceived quality. For these models the characteristics of the stream content and decoder strategies such as error concealment are usually taken into account. The model ITU-T J.247 [65] for video quality assessment belongs to this category.

- Parametric packet-layer model use only the packet header (TCP, RTP, UDP, IP, etc.) information without having access to the media signal. Such models are well suited for in-service non-intrusive multimedia quality monitoring. Among this category we may indicate the Recommendation ITU P.1201 [66].

- Parametric planning models use the quality planning parameters (bandwidth, packet loss rate, delay, frame rate, resolution, etc.) for network and terminals to predict the quality. For example, the models G.1070 [67] and G.1071 [68] are parametric models for estimating video and audio qualities for videotelephony and streaming applications respectively. The E-model (Rec. G.107) is a planning model for audio quality.

- Bitstream-layer models predict the QoE based on both encoded bit stream and packet-layer information without performing a complete decoding. These models can be used in situations where one does not have access to decoded video sequences. The Recommendations ITU P.1202[69] and P.1203 [70] are bitstream layer models for video and audiovisual media streaming quality assessment.

- Hybrid models are a combination of two or more models from the preceding. These models analyze the media signal, the bitstream information and packet header to estimate the perceived quality. For instance, ITU J.343 [71]is on of the developed hybrid models.

In the following sections we will describe in more details examples of models and metrics for audiovisual, video and audio quality assessment belonging to the above categories.

2.4.1 Audiovisual global quality metrics

2.4.1.1 ITU-T P.1201 model

P.1201 [66] describes a parametric non-intrusive model for the assessment of audiovisual media streaming quality. It is a no-reference algorithm for monitoring the audio, video and audiovisual quality of streaming based services. The model is composed of two sub-standards describing individual models for two types of application areas:

- ITU-T P.1201.1 specifies the model algorithm for the lower resolution (LR) application area (mobile TV)

- ITU-T P.1201.2 specifies the model algorithm for the higher resolution (HR) application area (IPTV).

The two ITU-T P.1201 model algorithms are no-reference (i.e., non-intrusive) models which operate by analyzing packet header information available from respective packet trace data, provided to the model algorithms in the packet capture format (PCAP). Further input information on more general aspects of the stream, such as the video resolution, which may not be available from packet header information, is provided to the model algorithm out-of band, for example in the form of stream-specific side information. As output, the model algorithms provide individual estimates of audio, video and audiovisual quality in terms of the five-point absolute category rating (ACR) mean opinion score (MOS) scale. Further, diagnostic information on causes of quality degradation can also be made available [66].

For mobile application area, the block diagram of the P.1201.1 model is shown in Fig 2.5.

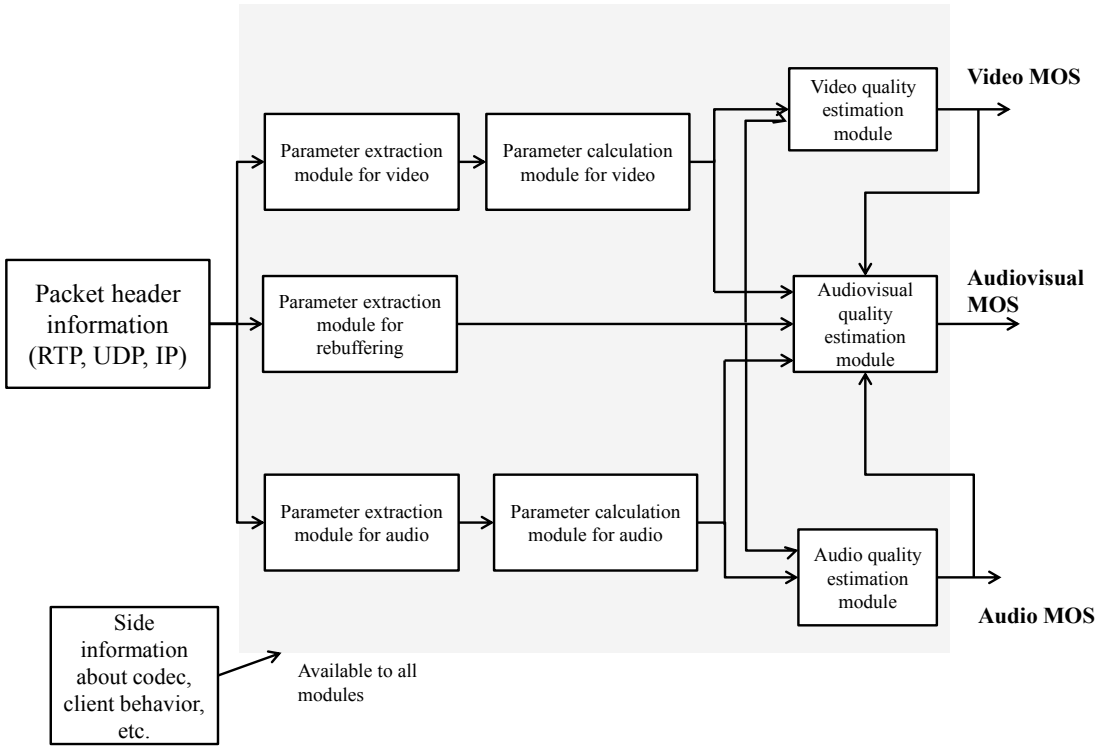


Figure 2.5: Block diagram of P.1201.1 model

Audio quality estimation module: Taking as input the packet header information, the RTP timestamp, sequence number, and payload parameters are extracted by the audio parameter extraction module. Then, the parameter calculation module estimates the length of the lost audio frame per audio RTP packet using the

extracted audio RTP timestamp and clock rate. At the same time, it calculates the number of audio packets per RTP timestamp. The lost audio frame length in milliseconds is the calculated using average audio burst packet loss length and audio frame length.

Video quality estimation module: Taking as input the packet header information, the video RTP timestamp, sequence number, market bit, and payload are firstly extracted. Then, the parameter calculation module for video estimates video packet-loss length based on the video RTP sequence number and lost bytes for lost video RTP packets using the same method as that of parameter calculation for audio module.

Audiovisual quality estimation module: is an integration model that combine scores calculated by video and audio modules into a global audiovisual quality score.

Recommendation ITU-T P.1201 is verified and recommended for unreliable content transmission (transmission over RTP/UDP for lower resolution, and transmission over MPEG2-TS/RTP/UDP or MPEG2-TS/UDP for higher resolution. Recently, a new Recommendation ITU-T P.1203 is published and restricted to reliable content transmission as in TCP protocols. ITU-T P.1203 is a parametric bitstream-based quality assessment model of progressive download and adaptive audiovisual streaming services over reliable transport [70].

2.4.1.2 ITU-T G.1070 model

ITU-T G.1070 [67] describes a parametric computational model for point-to-point videophone applications over IP networks standardized by ITU in 2012. The algorithm estimates the perceived quality based on measurement parameters, but not based on the actual video and audio signals. The inputs of the model are information about codec, coded bitrate, transport errors and client information about buffering.

The algorithm is trained to estimate the quality for typical and average audiovisual content, and give the same score for a given codec, bit rate and transport error situation independent of the audiovisual content.

This parametric algorithm is able to score live video, since detailed information about the source video is not required. The algorithm typically requires information about codec and coded bit rate. This type of algorithm may still be applicable when only an encrypted bitstream is available.

G.1070 model is composed of three quality modules: the Audio, the Video and the Audiovisual modules. As output, the modules provide individual estimation of the audio and video qualities and the model combines all of them in an integration function for overall audiovisual quality on the 5-point ACR scale (MOS_A, MOS_V and MOS_{AV}) [72].

The speech quality model is inspired by the recommendations G.107 and G.107.1 with the same parameters. The model is described as follows: For narrowband

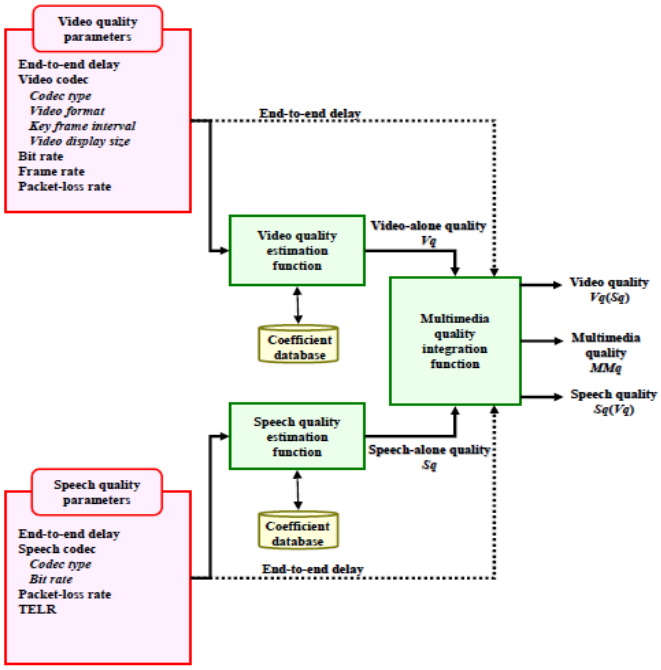


Figure 2.6: ITU-T Rec.1070 model

speech:

$$Q = 93,193 - Idte - Ie_{eff} \tag{2.1}$$

Idte defines the degradation caused by talker echo as follows:

$$Idte = [\frac{94.769 - Re}{2} + \sqrt{\frac{(94.769 - Re)^2}{4} + 100} - 1](1 - e^{-T_s}) \tag{2.2}$$

where

$$Re = 80 + 2.5(TERV - 14) \tag{2.3}$$

and

$$TERV = TELR - 40log\frac{1 + \frac{T_S}{10}}{1 + \frac{T_S}{150}} + 6e^{60.3T_S^2} \tag{2.4}$$

$Ie - eff$ represents the degradation caused by speech coding and packet loss and is defined as:

$$Ie_{eff} = Ie_s + (95 - Ie_s).\frac{Ppl_s}{Ppl_s + Bpl_s} \tag{2.5}$$

and for the Wideband audio

$$Q = 129 - Idte_{WB} - Ie_{eff,WB} \tag{2.6}$$

The video quality estimation module is defined by the following equation:

$$V_q = 1 + I_{coding}exp(-\frac{Pplv}{D_{Pplv}}) \tag{2.7}$$

where I_{coding} expresses the video quality impacted by the coding distortion: video bit rate and video frame rate. D_{Pplv} is the packet loss robustness factor which represents the degree of video quality robustness due to packet loss. $Pplv$ represents the packet loss rate.

The general audiovisual quality estimation module is:

$$MM_q = m_1 MM_{SV} + m_2 MM_T + m_3 MM_{SV} MM_T + m_4 \tag{2.8}$$

where MM_{SV} represents audio-visual quality and is defined as:

$$MM_{SV} = m_5 S_q + m_6 V_q + m_7 S_q V_q + m_8 \tag{2.9}$$

MM_T represents the degree of the audio-visual quality due to audio/video delay and synchronization and is expressed as:

$$MM_T = max(AD + MS, 1) \tag{2.10}$$

$$AD = m_9(T_S + T_V) + m_{10} \tag{2.11}$$

$$MS = min(m_{11}(T_S - T_V) + m_{12}, 0) if T_S \geq T_V \tag{2.12}$$

$$MS = min(m_{13}(T_V - T_S) + m_{14}, 0) if T_S < T_V \tag{2.13}$$

where AD is the absolute audiovisual delay and MS is the audiovisual media synchronization.

The G.1070 model can be applicable only for the conditions summarized in the table 2.1.

Codec type	MPEG-4, MPEG-3, ITU-T G.1070
Video format	QVGA, QQVGA, VGA
Video display size (inch)	4.2, 2.1, 9.2

Table 2.1: Conditions of ITU G.1070 model

In the literature, various studies suggested methods to improve the quality assessment accuracy of the G.1070 model and thus to get better correlation between the calculated objective scores and the results of subjective tests. Considering that the video quality have high variations depending on video content, in [73, 74] Joskowicz et al. proposed an enhancement of the model by taking into account the characteristics of the video content. A new parameter representing an estimation of spatial-temporal activity is included in the model. The evaluation of the enhanced model shows that it performs much better than the original model and correlates better with the subjective quality perception.

In [75] Narvekar et al. presented a method to estimate the video parameters of the G.1070's input in order to use the G.1070 video quality estimation model for monitoring applications. An estimation function is added to compute bit rate, frame rate, and packet loss rate from the received encoded video bit stream. These parameters are then used by a G.1070 video quality model.

In order to make the G.1070 model up to date with the continuous development of the video telephony applications, subjective studies are conducted to propose a set of coefficients to extend the G.1070 opinion model to support current generation of video codecs (H.265/HEVC, VP9) and full-HD video format [76, 77].

More recently, Huawei Technologies Co. Ltd. are interested in studying the G.1070 model and are conducting studies to extend the usage cases of the model. The video codecs type are in continuous development and the resulted stream quality is influenced. Thus, Jing Xiao and Shijun Zhang carried out a series of training experiments to obtain coefficients for the H.264 codec in its Hight Profile (HP) and Baseline Profile (BP) with different parameters. On the basis of eight tests, parameter values for the parameters v1 to v12 have been derived for the H.264 codec on small screen (6 inches). Details on the tests are described in the contributions C-129 [78] and C-130 [79] discussed in the ITU meeting on September 2017. The results have been fitted using the algorithm described in Rec. G.1070.

Codec	H.264 BP	H.264 BP	H.264 BP	H.264 BP	H.264 HP	H.264 hP	H.264 HP	H.264 HP
Format	VGA	4 CIF	720p	1080p	VGA	4 CIF	720p	1080p
Bit rate (bps)	128, 192, 512, 768, 1024	128, 256, 512, 1024, 1280	256, 384, 512, 2048, 3200	512, 768, 1024, 4096, 6400	128, 192, 512, 768, 1024	128, 256, 512, 1024, 1280	256, 384, 512, 2048, 3200	512, 768, 1024, 4096, 6400
Frame rate (fps)	8, 15, 30	8, 15, 30	8, 15, 30	8, 15, 30	8, 15, 30	8, 15, 30	8, 15, 30	8, 15, 30
Packet loss rate (%)	0,0.5, 1,3	0,0.5, 1,3	0,0.5, 1,3	0,0.5, 1,3	0,0.5, 1,3	0,0.5, 1,3	0,0.5, 1,3	0,0.5, 1,3

Table 2.2: Conditions of ITU G.1070 extended model

2.4.2 Video quality metrics

The objective evaluation of video quality was first performed with simple signal processing tools such as the Peak signal-to-noise ratio (PSNR) and Structural SIMilarity index (SSIM). These metrics are basically used for fixed image quality evaluation. In order to adapt these image based metrics, extended versions are developed applying temporal pooling methods. Various research studies have shown that these objective measures are limited and do not correlate well with subjective opinions [80, 81]. The use of more elaborated objective methods is therefore necessary. Then, the major challenge is to design a metric that models the behavior of the human visual system. This problematic has already been studied on a large scale and a variety of algorithms for video quality estimation have been proposed [82, 83]. However, few standards are developped and confusions on representative quality metrics exist.

The objective video quality metric models that result in ITU are mainly validated in the Video Quality Experts Group (VQEG)[84].

Some researchs proposed parametric-packet layer, parametric planning, and bitstream-layer models in [85, 86]. Commonly, the limitation of the packet-layer and bitstream-layer models is the fact that they are adapted to specific codecs and network protocols. Furthermore, they are dedicated to particular services, in general for IPTV and streaming video quality assessment.

In this thesis we are interested in full reference media-layer video quality models and no-reference media-layer video quality metrics. We propose an up-to-date review and performance comparison of the existing metrics.

2.4.2.1 Full Reference (FR) metrics

Research studies have been conducted to compare the performance of the state-of-the-art full reference objective video quality metrics. In one of the most recent reviews (2011) S. Chikkerur et al. [87] classified full reference metrics into three categories: (i) traditional point-based metrics (MSE, PSNR), (ii) natural visual characteristics and (iii) perceptual Human Visual System (HVS). They performed a comparison evaluation of the reviewed FR metrics on the LIVE Video quality database. They found that the metrics MS-SSIM, VQM and MOVIE are the best performing video quality assessment algorithms. Further, principles of perceptual models for predicting video quality and survey of objective metrics are investigated in [82, 88, 89].

In video quality assessment, Full Reference (FR) metrics perform a comparison between a reference free degradation video stream and a distorted video stream. In this type of approach, we assume that the loss of quality is directly related to an error signal added to a signal initially "Perfect". Since this type of metrics requires the entire reference video to be available, they are not useful in real time evaluation and in monitoring. Full Reference metrics generally impose a precise spatial and temporal alignment of the two signals.

The selected algorithms, later studied, are widely cited in the literature, and have been reported to have good performance. Moreover, the authors of the selected metrics have released the source codes of their respective metrics. Therefore, the presented results are easy to reproduce. The ten FR video quality assessment metrics described in the following subsections include Peak Signal to Noise Ratio (PSNR), Structural SIMilarity index (SSIM) [105], Multi-Scale Structural SIMilarity index (MS-SSIM) [92], Video Quality Metric (VQM) [93] (including its general model and videoconferencing model), MOtion-based Video Integrity Evaluation (MOVIE) [95] , ViS3 [97], SSIMplus [99] and Video Multi-method Assessment Fusion (VMAF) [101].

Objective MOS prediction metrics are also standardized by the ITU (J. series recommendations) to assess the video quality. It would be interesting to compare

Metric	**Year**	**Approach**	**Pooling method**	**Value Range**	**Execution time** (normalized based on PSNR)	**Tool**
PSNR		Mean square error measurement	Mean over the frames	[0, 100]	1	MSU software [90]
SSIM [91]	2004	Structural distortion measurement	Mean over the frames	[0, 1]	1.05	MSU software [90]
MS-SSIM [92]	2003	Multi-scale structural distortion measurement	Mean over the frames	[0, 1]	2	MSU software [90]
VQM [93]	2004	Edge impairment filter	Compute Temporal Information (TI)	[0, 1]	30	NTIA software [94]
MOVIE [95]	2010	Gabor filter bank	Temporal distortions index	[0, 1]	456	Source Code [96]
ViS3 [97]	2014	detection-based and appearance based strategies of the MAD algorithm	Spatiotemporal dissimilarity index	[0, 100]	23	Matlab code [98]
SSIMplus [99]	2015	Contrast sensitivity function	Mean over the frames	[0, 100]	4	SSIMwave software [100]
VMAF [101]	2016	Machine Learning	Temporal information among the elementary metrics	[0, 100]	26	Source code [102]
OPVQ [103]	2016	ITU-T J.247	Mean over the frames	[1,5]	19	OpenVQ Toolkit [104]

Table 2.3: Characteristics of full reference objective metrics

their prediction accuracy with the diverse full reference metrics. Unfortunetly, we do not have access to these models because of their commercial licenses. For instance, the model J.247 is owned by the company OPTICOM. We introduce in our study an open source implementation of this model named OPVQ [103]. In Table 2.3 we summarize the characteristics of the surveyed metrics. In the following subsections, F_{ref} and F_{dist} denote the reference and distorted video frames respectively. The subscript ref denotes reference and $dist$ distorted video streams. Moreover, W and H represent the width and the height of videos respectively.

Peak Signal to Noise Ratio (PSNR)

PSNR is the most widely used FR objective signal distortion and quality metric. It is a pixel base signal quality comparison metric by quantifying the error between the distorted signal and the reference signal and is defined as:

$$PSNR = 10 \log_{10} \frac{L^2}{MSE} \tag{2.14}$$

Where L is the dynamic range of the pixel values, e.g., for 8 bits/pixel image we have $L = 2^8 - 1 = 255$ and MSE is the Means Square Error defined as:

$$MSE = \frac{1}{WH} \sum_{j=1}^{H} \sum_{i=1}^{W} (F_{ref}(i,j) - F_{dist}(i,j))^2 \tag{2.15}$$

The PSNR is used to express the quality of reconstruction of an image compression lossy algorithm. When the reference and the degraded images are identical, the value of PSNR is undefined ($+\infty$). It is very commonly used because there are many situations where its use makes sense and is very suitable for optimization methods. Moreover, its simplicity calculation and execution speed are arguments that justify its use quasi-exclusive by the signal processing community. Furthermore, there is currently little metric questioning its use. PSNR is highly criticized because it does not well correlate with the human perception of the measured quality. Indeed, it does not model the human visual system, assumes that the visual quality decreases when signal distortion increases. However, it is well known that the quality depends not only on distortions but also on the content of the image, or also on the location of distortions.

Moreover, in the case of video assessment, approaches such as PSNR, do not take into account the temporal content of the video as they are calculated on each image pixel by pixel, which sometimes has a disastrous effect on metric results (time-synchronization, spatial or temporal misalignment).

Finally, across contents there is no strong and consistent relationship between these metrics and the average subjective opinion score of observers. Researches in recent decades tend to develop objective metrics, essentially full reference, taking into account the characteristics of the human visual system. Other approaches, such as structural approaches have been implemented based on local similarities.

Structural Similarity (SSIM)

It is developed by Z. Wang et al. and presented in [105]. SSIM is a metric that calculates the similarity between two signals. Basically, it is used for quality assessment of images. In the case of video signals, SSIM index is applied frame-by-frame on the luminance component of the video [91], and the overall SSIM index for the video is computed as the average of the frame-level quality scores. Unlike PSNR, it does not compare images pixel by pixel but by properly selected small $N \times N$ blocks. Thus, SSIM is sensitive to the structural distortions as the case of the human eye sensitive to changes in the structure. Consequently, SSIM has a significantly reduced computational costs and still provide good experimental results.

Similarity index is measured within sliding window. Thus, the formula to calculate SSIM between two windows X and Y of common size $N \times N$ is:

$$SSIM(F_{ref}, F_{dist}) = \frac{(2\mu_{ref}\mu_{dist} + c_1)(2\sigma_{ref,dist} + c_2)}{(\mu_{ref}^2 + \mu_{dist}^2 + c_1)(\sigma_{ref}^2 + \sigma_{dist}^2 + c_2)} \tag{2.16}$$

With μ_{ref} and μ_{dist} are the average intensities of F_{ref} and F_{dist} respectively. σ_{ref}^2 and σ_{dist}^2 are the variances of F_{ref} and F_{dist} respectively, $\sigma_{ref,dist}$ is the covariance of F_{ref} and F_{dist}. c_1 and c_2 are two variables to stabilize the division with weak denominator.

Multi-scale Structural Similarity (MS-SSIM)

MS-SSIM is an extension of SSIM index that incorporates the details of the frame at different resolutions (or scales) [92]. A low-pass filter is iteratively applied to the reference and degraded frame. Then, a process of sub-sampling of the filtered image by a factor of 2 from the previous iteration is applied. At each scale, the MS-SSIM algorithm evaluates the value of SSIM and attributes less weight to the luminance term unlike the contrast and structure terms.

MS-SSIM is computed as fellow:

$$MS_{SSIM} = \frac{1}{M_w} \sum_{i=1}^{M_w} SSIM(F_{ref}^i, F_{dist}^i) \tag{2.17}$$

With M_w is the total number of scales, and F_{ref}^i and F_{dist}^i are frame contents at the $i-th$ local window. We used the extension of the MS_{SSIM} index to video by applying it frame-by-frame on the luminance component of the video and the overall MS_{SSIM} index for the video was computed as the average of the frame level quality scores. This metric has been shown to outperform the SSIM index and many other image quality assessment algorithms.

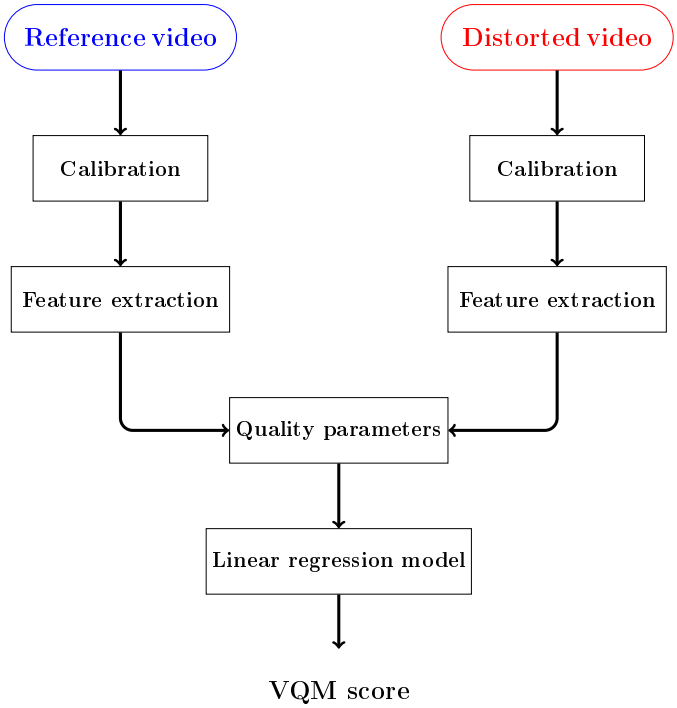


Figure 2.7: Block diagram of the VQM-G general model [1]

NTIA General Model (VQM-G)

VQM NTIA metric [1] proceeds in several steps as shown on Fig. 2.7.

A calibration step is performed to compare the sequences to be evaluated, considering:

- Alignment and spatial adjustment between the two sequences: specify the horizontal and vertical spatial shift of the processed video relative to the original video.
- Estimation of the region of interest on both the original and processed video streams for feature extraction: a column of pixels "J" does not belong to the region of interest if it is black (mean of pixels values $M_J < 20$) or if the average pixel level of the mean value for successive columns indicates a black border ($M_{J-1} > M_J$).
- Estimation of the perceived contrast and brightness level.
- Alignment and temporal correction between the two sequences: estimating video delay by correlating lower resolution frames, sub-sampled in space and extracted from the reference and degraded video streams.

The calibration makes the VQM metric not sensitive to horizontal and vertical shifts of the image, temporal shifts of the video stream, and changes in image contrast and brightness.

Then, local features are extracted from the reference sequence and the distorted one before comparing them. The extraction of the features is performed by elementary spatio-temporal regions. Such a region is a set of pixels defined by its two spatial dimensions and its temporal dimension. The VQM model computes six features. Two features characterize spatial activity: derived from horizontal and vertical spatial gradients to describe perceptual distortions of edges (blurring and blocking). The third feature characterizes distortions in chromatic components, the fourth characterizes local contrast. The fifth feature represents the amount of temporal information (the standard deviation of the absolute value of the difference between consecutive video frames at time t and t-1 and). Finally, the sixth represents the product of the features of the local contrast and the Temporal Information (TI).

By comparing the extracted features from the processed video with those extracted from the reference video, quality parameters that describe changes in the video quality are computed. Three comparison functions are used: error ratio, logarithmic ratio, and the Euclidean distance. Finally, a linear regression of these parameters defines the global VQM measure.

NTIA Videoconferencing model (VQM-V)

This model is optimized to achieve maximum objective to subjective correlation for videoconferencing context [106]. The difference between this model and the general model described above is the selection of the used parameters. Videoconferencing model consists of a linear combination of six parameters. Four parameters are based on features extracted from spatial gradients of the Y luminance component, and two parameters are based on features extracted from the absolute temporal information of the Y luminance component. The impairment types measured by VQM-V model are blurring, block distortion and jerky/unnatural motion. Error blocks and color distortions included in the general model are not present in VQM-V model.

Open Perceptual Video Quality metric (OPVQ)

OPVQ is an implementation of the model described in ITU-T J.247 Annex B [103]. The algorithm has four main steps presented in the block diagram in Fig. 2.8.

The first step is a simple pre-processing step consisting of some predefined cropping based on the video resolution. Next, fine alignment is done in the spatial domain, i.e. the sequences should at this point be aligned from start to finish. Chroma correction is also performed, using histogram correction. The third step is the distortion analysis which generates four separate indicators. The first two ones measure intra-frame distortion for the luma and chroma channels respectively. Distortion is measured as introduction or loss of edges in a specific frame. Indicators three and four measure inter-frame distortion, i.e. the amount of change at a specific

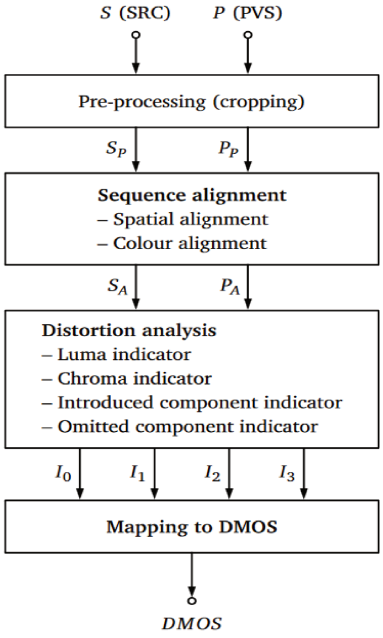


Figure 2.8: The block diagram of the OPVQ algorithm

position between two adjacent frames. The indicators are based on the difference of the components at the same spatial position in the corresponding pair of adjacent frames from reference and distorted sequences. At the fourth and last step, these indicators are weighted using parameters specific to the resolution, and mapped to a single mean opinion score (MOS). OPVQ as described in J.247 provides support for only a limited set of spatial resolutions (VGA, CIF and QCIF).

MOtion-based Video Integrity Evaluation (MOVIE)

It is a full-reference video quality index developed at the Laboratory for Image and Video Engineering (LIVE) by K. Seshadrinathan and Bovik [95]. The metric is based on models of human vision system and consists of two indexs: Spatial MOVIE index that captures spatial distortions and Temporal MOVIE index that captures temporal distortions.

The temporal component of MOVIE uses optical-flow motion estimation to determine motion information from the reference video, which is combined with the outputs of the spatio-temporal Gabor filters (3 scales, 35 filters at each scale) to capture temporal distortion.

$$TemporalMOVIE = \sqrt{\frac{1}{\tau}\sum_{j=1}^{\tau} FQ_T(t_j)} \tag{2.18}$$

With FQ_T is the frame level quality index for temporal MOVIE. It is defined as

the standard deviation to the mean of the Temporal MOVIE scores for that frame.

The spatial component employs the outputs of the spatio-temporal Gabor filters applied on a multi-scale decomposition of the reference and the distorted videos. Then, a model of contrast masking captures the spatial distortions. Spatial distortions in the video such as blur, ringing, false contouring, blocking, noise and so on can be captured using errors computed between corresponding Gabor sub-bands of the reference and test videos.

$$SpatialMOVIE = \frac{1}{\tau}\sum_{j=1}^{\tau} FQ_S(t_j) \tag{2.19}$$

FQ_S is the similar as defined for the Temporal MOVIE. The final MOVIE index for the video sequence is computed as the product of these two index.

$$MOVIE = SpatialMOVIE \times TemporalMOVIE \tag{2.20}$$

The key difference of this method is that a subset of spatio-temporal Gabor filters are selected adaptively at each location based on the direction and speed of motion, such that the major axis of the filter set is oriented along the motion trajectories of the reference video. The video quality assessment process is carried out with coefficients computed from these selected filters only.

Vis3

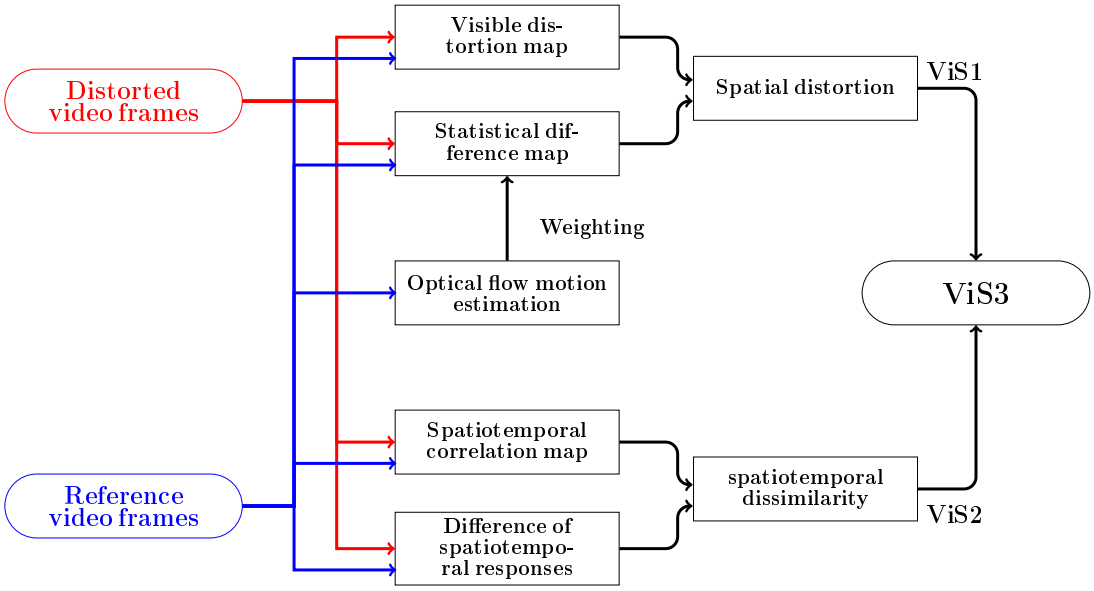


Figure 2.9: Vis3 diagram chart

It was recently proposed by Phong V. and Damon M. Chandler in [97]. The algorithm estimates video quality by measuring spatial distortion (ViS_1) and spatio-temporal dissimilarity (ViS_2).

Spatial distortion is estimated by applying a detection-based strategy (computes the perceived distortion due to visual detection) and an appearance based strategy (computes the perceived distortion due to visual appearance changes) of the Most Apparent Distortion (MAD) algorithm [107] to Groups of video Frames (GOF). For each group of consecutive frames, a visible distortion map is computed by using MAD's detection-based strategy. Both the reference and the distorted frames are converted to perceived luminance and filtered by a contrast sensitivity function. A local distortion visibility map is obtained by comparing the local contrast of the reference frame and the distorted frame. This map is then weighted by local mean squared error to yield a visible distortion map.

Then, a statistical difference map is computed by using MAD's appearance-based strategy. The reference and the distorted frames are decomposed into different subbands using a 2-D log-Gabor filter-bank. Local standard deviation, skewness, and kurtosis are computed for each subband of both the reference and the distorted frames. The differences of local standard deviation, skewness, and kurtosis between each subband of the reference frame and the respective subband of the distorted frame are combined into a statistical difference map.

The effect of motion on the visibility of distortion is modeled using the optical flow motion estimator. In fact, greater weights are given to spatial distortions in the slow moving regions. These per-group maps values are then combined into a single spatial distortion map. The spatial distortion value is defined as the root mean square (RMS) value of this map.

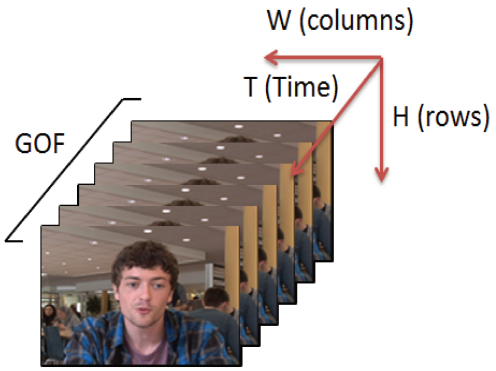


Figure 2.10: Spatio-Temporal Slices (STS)

Spatio-temporal dissimilarity estimates video quality degradation by computing the differences of spatio-temporal responses of modeled visual neurons from the reference and distorted videos. This index is computed via the use of Spatio-Temporal Slice (STS) images as shown on Fig. 2.10 . First, an extraction of the vertical and horizontal STS frames in the luminance component is performed. Then, spatiotemporal correlation map of the STS frames (local linear correlation coefficients) and the difference of spatiotemporal responses are computed in a block-based fashion and combined to yield a spatiotemporal dissimilarity map. All maps are then collapsed by using root mean square and combined to yield the spatiotemporal dissimilarity

value ViS_2 of the distorted video.

The final estimated score of the perceived video quality degradation is a geometric mean of the spatial distortion and the spatiotemporal dissimilarity values.

$$ViS_3 = \sqrt{ViS_1 \times ViS_2} \tag{2.21}$$

SSIMplus

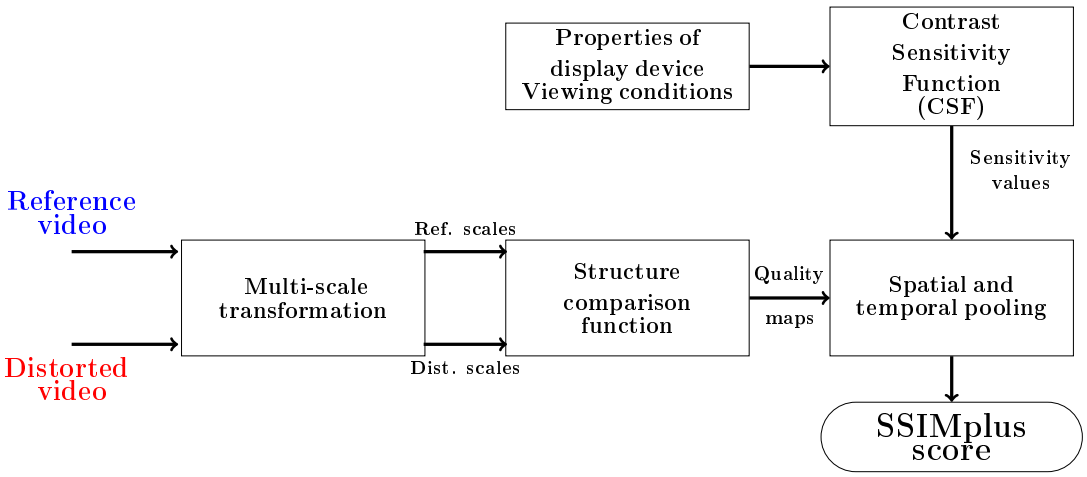


Figure 2.11: SSIMplus diagram chart

SSIMplus is among the latest metric that has been developed by Zhou Wang's team [99]. This objective metric evaluates the quality of experience QoE for video services unlike the other metrics (SSIM, MS-SSIM ...) that assess pure video quality without taking into account the conditions and the viewing context. For example a video that has VGA resolution and looks great on an iPhone might look awful when it is displayed in full screen on a 31" monitor or 55" TV. Here we could evocate the limitations of video quality assessment (PSNR, SSIM ...)

- Network condition not considered.
- Receiving device (speed, power, memory...) not considered.
- Display device not considered.
- Display resolution not considered.
- Viewing condition/environment not considered.

SSIMplus algorithm offers the ability to apply the metric to different viewing devices and conditions. The available display devices considered in calculating QoE scores are: iPhone 5S, iPad Air, Lenovo W530 laptop, Sony 55" TV, Sony 55" TV (TV-Expert). SSIMplus rates the videos on a scale of 1 to 100, with 20-point gaps

separating the video into meaningful human measure: bad (0), poor, fair, good, or excellent (100) as it looks to a human viewer.

The SSIMplus algorithm performs a multi-scale transformation on the reference and distorted video frames (see Fig. 2.11). Then, a quality maps are computed based on a structure comparison function between subsequent reference and distorted scales. The quality of all the scales is determined by performing spatial pooling of the quality maps based on the local information content and distortion. The perceptual quality of the distorted frame is calculated using a weighted combination of the scale-wise quality values. The weights are determined using a method that takes into account the properties of the display device and viewing conditions. These parameters include: 1) average or range of user viewing distance, 2) sizes of viewing window and screen; 3) screen resolution; 4) video scaling; 5) screen contrast; 6) replay temporal resolution; 7) illumination condition of the viewing environment; 8) viewing angle; 9) viewing window resolution; 10) post-filtering and image resizing methods; 11) device model; 12) screen gamma correction parameter; 13) video scan type (interlaced or progressive).

Video Multi-Method Assessment Fusion (VMAF)

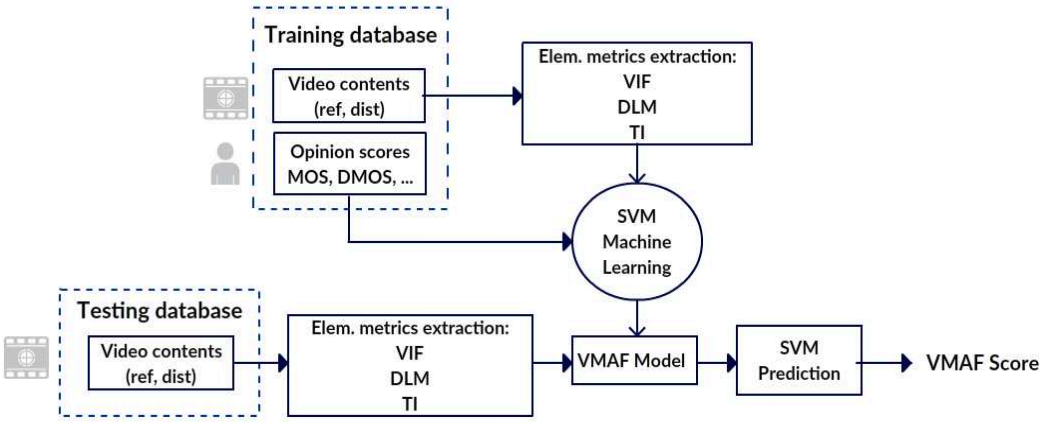


Figure 2.12: The block diagram of the Vmaf model

It is a recent metric developed by a research group of the University of South California and the company Netflix and published in June 2016 [101]. For the purposes of improving its video streaming service with automated quality monitoring, Netflix is interested in the quality of the videos broadcast. According to the authors, the main properties that must satisfy a video quality metric are: 1) accuracy in capturing human perception of quality, 2) consistency across contents, 3) possibility to be run at scale and 4) adequate to streaming use case. Netflix proposes a full reference predictive model based on a machine learning regression algorithm, specifically the Support Vector Machine (SVM) algorithm. As its name suggests, the VMAF predict the video quality by combining multiple elementary quality metrics.

- Visual Information Fidelity [108] is a statistical visual image quality model in wavelet domain. It is based on distortion and Human Visual System (HVS) modeling. VIF score is the ratio between the visual information of the distorted and the reference frame. The visual information of both frames is quantified by the mutual information between the input frame and the output of the HVS model.

- Detail Loss Metric (DLM)[109] is an image quality assessment algorithm. It refers to measuring the loss of useful information which affects the content visibility.

- Motion metric quantifies the amount of motion in a sequence. It is measured by the temporal difference between adjacent frames.

Since the VMAF model is constructed from a machine learning algorithm, the prediction performance and accuracy depend on the consistency of the learning database. Thus, Netflix has generated a video data set that reflects the types of artifacts that cause the degradation of video streaming quality. Two types of impairments are identified: compression artifacts and scaling artifacts. Thus, the model is trained on 334 sequences (34 references and 300 distorted) of 6 seconds long with different contents (TV shows and movies) and different characteristics in terms of spatial and temporal complexities. The source videos are encoded with H.264/AVC at wide range of bitrates from 375 kbps to 20.000 kbps and at resolutions ranging from 384×288 to 1920×1080.

2.4.2.2 No Reference (NR) metrics

Most recent researches developed accurate full reference models that correlate well with the human video quality perception. However, these metrics are not useful for monitoring and troubleshooting an application working in real time as this is the case for video conference calls. In this context, there is a great need to focus on no reference metrics.

Most of the no-reference approaches estimate video quality by qualifying the presence of some degradations in the video stream. The most commonly quoted indicators are thus linked with conventional impairments such as: bluriness, blockiness or jerkiness, known as the most common artifacts of compression methods (H.26x, MPEG and their derivates). A blocking measure for compression video sequences has been proposed by Vlachos in [110]. Eventually, several other implementations of the metric are developed to detect the block effect in videos. A performance comparison between three blockiness metrics has been carried on by S. Winkler et al. in [111].

X. Yuanyi et al. [112] developed a new no-reference video quality metric for detecting temporal jerkiness caused by frame freezing. Their algorithm is based on detecting freeze events, extracting a set of features corresponding to the distortion and then training a neural network on a large database. In real time transmission,

packet loss artifacts are the common source of quality degradation for conversational applications over IP and wireless network. This is not the case for adaptative streaming over TCP. In [113], authors developed a metric that evaluate the quality of sequences reconstructed after packet loss impairments. Based on the video stream, they analyze the continuity of macroblock data on edges between consecutive frames.

Other studies tried to combine a set of objective measures in order to generate a NR model that estimate the overall video quality [114, 115]. They trained a multilayer perception neural network (MLP) estimator of global MOS quality score. Working on the bit-stream, A. Raake et al. [116, 117] proposed a no-reference method for estimating the visibility of packet losses in standard (SD) and high definition (HD) H.264/AVC video sequences. Within the Study Group 12 of the International Telecommunication Union (ITU), parametric no reference standards (P.1201, P.1202 and P.1203) are developed [69, 66, 70]. These models estimate global video quality by analyzing the bitstream packet-header information.

2.4.3 Audio quality metrics

The objective measurements of voice quality in modern communication networks can be intrusive or non-intrusive. The intrusive methods analyze the transmitted (original) and received (degraded) speech signals. Thus, these methods compare the reference speech signal with the corresponding distorted signal. Non-intrusive methods allow estimation of the perceived voice quality by exploiting information extracted from the receiver side. Non-intrusive methods use only the degraded (received) speech signal to estimate the corresponding voice quality [63]. Intrusive methods are more accurate than non-intrusive ones, but they are not suitable for real-time traffic monitoring because of the need for reference data. A typical intrusive method is based on the latest ITU P.863, Perceptual Objective Listening Quality Assessment (POLQA) [14].

Non-intrusive methods are more appropriate for real-time traffic monitoring since they do not need the reference signal. There are two categories of non-intrusive methods: those based on the signal and those based on parameters-based method. An example of a non-intrusive signal-based method is the *vocal tract model* [118], which aims to predict voice quality by directly analyzing the speech signal being listened to (a degraded signal) without the reference signal.

Intrusive/signal-based methods: POLQA, ETSI TS 103 281

The basic idea of intrusive methods is that a signal is injected into the system under test, and the degraded output is compared by the objective test system to the input signal considered as the reference. Therefore, intrusive assessment techniques require access to both the transmission and reception ends of communication.

The following models are typically used:

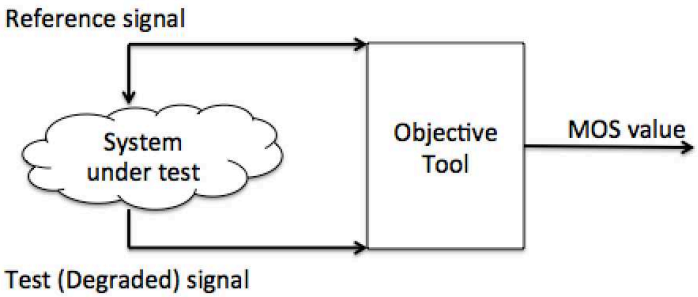


Figure 2.13: Intrusive objective tool implementation

POLQA: POLQA is the result of collaboration between three companies (Opticom, Swissqual and TNO) and was standardized by ITU-T in 2011 in the Recommendation P.863 [14]. POLQA takes into account signals in narrowband, wideband and super-wideband (50-14000 Hz). It can be used for the evaluation of speech transmission quality in 3G, 4G / LTE and VoIP networks, and speech processing systems such as noise reduction systems and so on.

POLQA only takes into account the impairments related to the listening context such as ambient noise at the speaker level, the loss of packets ... The impairments perceived during a conversational situation such as the echo are not taken account by this model. However, contrary to the objective models of the listening context, POLQA integrates a module estimating the impact of reverberation on the quality which is a phenomenon rather related to the context of phrase or conversation. In addition, the reference and input gradient signals of this model may be electrical or acoustic in nature (i.e. the signals are captured via an acoustic interface). It operates in two operating modes, one of which is dedicated exclusively to narrowband audio signals (NB mode) and the other allows application to audio signals up to super-wide band (SWB mode) and covers all three audio bands (narrow band, enlarged band and super-enlarged band). POLQA provides an overall quality score ranging from 1 to 4.5 for NB mode and from 1 to 4.75 for SWB mode.

ETSI TS 103 281 model: ETSI TS 103 281 [119] describes two models addressing the speech quality, background noise quality, and overall quality, as measured according to ITU-T Rec. P.835. It predicts the speech quality experienced with super-wideband and fullband terminals in the presence of background noise. The Technical Specification also provides evaluation results comparing model predictions to subjective data. Further, ETSI TS 103 106 [120] describes a model used with mobile terminals, as well as an evaluation of model performance. It can estimate quality in 3 dimensions: S-MOS-LQO (speech quality), N-MOS-LQO (noise intrusiveness), G-MOS-LQO (global quality).

Non-intrusive, parameter-based methods: E-model, P.564
For network planning, objective models have been developed to predict quality based on parameters.

Recommendation P.564: Recommendation P.564 [121] defines a set of minimum performance criteria to be achieved by single-end objective models in a listening context such as PsyVoIP [122] and VQMon respectively developed by Psytechnics and Telchemy. These models are mainly used to monitor the real-time transmission quality of IP networks. They estimate speech quality from the information contained in the Real-Time Protocol (RTP), User Datagram Protocol (UDP), and IP protocol headers, such as the packet loss rate, the type of codec used, and so on.

E-model: The E-model is a non-intrusive model for planning and predicting the voice quality of end-to-end transmission. It was developed by ETSI [123]) as an end-to-end tool for network designers and later standardized by the ITU in Recommendation G.107 [124]. The E model is used to measure echo, transmission delay and modern transmission impairments such as non-linear impairments related to low-rate codecs. Thus, it may be applied to predict voice quality in a conversational situation. The quality of transmission is expressed using a scalar called "transmission evaluation factor", noted R, whose expression is given by:

$$R = R_0 - I_s - I_d - I_{e,eff} + A \tag{2.22}$$

where

- R_0: basic signal-to-noise ratio (SNR), including noise sources such as circuit noise and room noise. It is the value that we obtain if the transmission is perfect.
- I_s: combination of all impairments, which occur more or less simultaneously with the voice signal.
- I_d: qualifies the impairments caused by delay and the echo.
- $I_{e,eff}$: impairments caused by low bit rate codecs and packet losses.
- A: allows the E-model to take into account the users indulgence toward the quality of the communication systems used (wired system, mobile, the terminal used, the use of the hands-free kit).

In the context of narrowband telephony, the scalar R values vary between 0 (very poor quality) and 100 (excellent quality). In addition, the factor can be converted to a MOS score (scale ranging from 1 to 5) as follows:

$$MOS_{CQE} = \begin{cases} 1 & \text{si } R < 0 \\ 1 + 0.035R + R(R-60)(100-R).7.10^{-6} & \text{si.} 0 < R < 100 \\ 4.5 & \text{si } R > 100 \end{cases} \tag{2.23}$$

where MOS_{CQE} is the estimation of voice quality in a conversational situation. A simplified version of the model has been proposed [125]. This version takes into account only the degradation caused by the codecs and the network conditions. Its expression is given by:

$$R = R_0 - I_{codec} - I_{packetloss} - I_{delay} \tag{2.24}$$

where the parameters I_{codec}, $I_{packetloss}$ and I_{delay} quantify the defects introduced by the codecs, the packet losses and the transmission delay. Model E was primarily for narrow-band telephony communications until 2011. Its extension to Wideband transmissions is standardized in ITU-T Recommendation G.107.1 where the maximum value of the R factor is 129 [126].

E-model has been a key element for evaluating the performances of different network for various telecommunication services. We found in [127] a review for some evaluations of E-model. Some modified E-model examples were presented in [127] to be more suitable for VoIP service. The applicability of E-model in the case of VoLTE was discussed and the necessity of studying jitter buffer algorithms was considered.

Chapter 3

Subjective tests and databases: experimental results

Contents

Introduction

Measuring the audiovisual quality of a multimedia stream is a complex task. Today, there is no other evaluator than the human eye and ear, coupled with his brain. This is why it is interesting to involve human observers and to ask them for their qualitative judgment in the evaluation of a videoconferencing service. However, carrying out such subjective tests raises many questions when they are implemented. The variability of human judgment, the control of the conditions of evaluation and the number of judgments necessary for a given representative judgment are all parameters to master.

As we saw in Chapter 2, the majority of these points have been the subject of research, and sometimes standardized at the international level mostly by the ITU. These standardized methodologies are different from each other and can be adapted to the context of the subjective test and to the application. In this chapter we present the methodologies and the processes of the subjective tests we implemented. Then, we introduce some statistical tools useful to the analysis of the results provided by the methodologies.

Our subjective studies have two purposes: to assess the perception of video conference service users under different conditions, and to constitute a sequences database to evaluate the performance of the objective quality metrics. We investigate the video, audio and audiovisual quality and asynchrony perception under two different situations: a non-interactive and an interactive conversational one. We analyze the effects of network impairments (packet loss, delay) on perceived audiovisual, audio and video quality. We evaluate the impact of experimental context and scene complexity on the quality perception in case of video calls. Furthermore, we propose new acceptability thresholds of audio-video asynchrony in video telephony context and study the effect of synchronization in the presence and absence of network degradation. The audio/video synchronization perception is more investigated in a specified study that we present in Chapter 4.

The remainder of this chapter is organized as follows. Section 3.1 defines the common elements of our test procedures. The statistical analysis that we performed are explained in Section 3.2. The experimental results of the conducted subjective tests are presented and discussed in Section 3.3 and 3.4. Section 3.5 introduces the external databases we used to complete our subjective test database in order to evaluate the objective models.

3.1 Common elements of the test procedures

The implementation of a subjective audiovisual quality test must comply with the recommendations of the ITU to ensure the reliability and reproducibility of the test. Although they are intended for different measurements, the standardized methodologies that we present share some common experimental conditions. These conditions are the panel of observers, the environment of observation and the global conduct of the sessions.

3.1.1 Selection of subjects

We know that for the same observed sequence, the judgments given by different individuals are generally not identical. In other words, evaluation is not stable from one individual to another. Several factors are responsible for this, such as the state of fatigue, knowledge of the sequence, the observer's general experience in video quality assessment, or personal appreciation.

In our subjective tests performed in the laboratory we call on non-expert observers, i.e they are not confronted with the video and audio quality evaluation in their professional activity. All participants are examined for their visual acuity through the Snellen test (Figure 3.1) and their color perception defects through the Ishihara test (Figure 3.2). The observer should have a visual acuity of 10/10 for both eyes with or without correction. Moreover, we made sure that all the subjects reported having a normal audition. For greater reliability of the results, a panel between 15 and 20 of participants will give statistically usable results [61, 48]. The panel should also be representative in age, gender and experience. We recruited the subjects from a "Testers database" of Orange and we paid them to participate in the experiments.

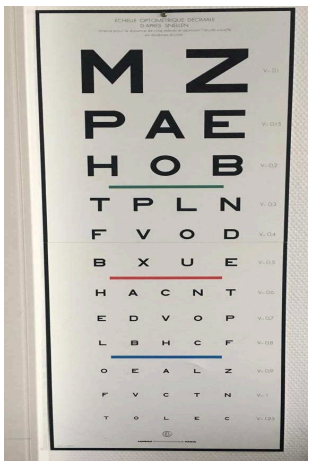


Figure 3.1: Used visual acuity test (Snellen).

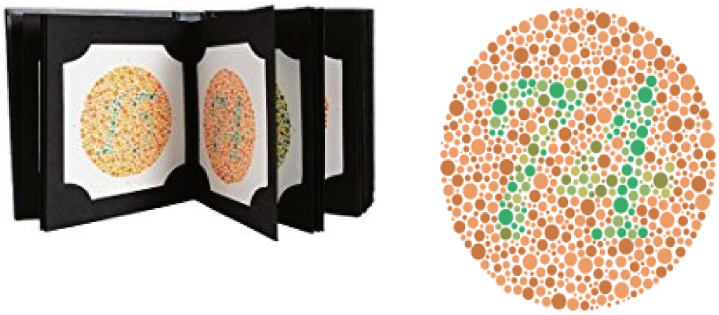


Figure 3.2: Color test (Ishihara).

3.1.2 Laboratory and test environment

Normalization of the visualization environment reduces the influence of the outside world on the observation and the evaluation of the sequences. The Recommendation ITU-T P.910 contains a number of rules to standardize the test environment. The three commonly measured factors that must be considered for the test environment are: the general environment (brightness, ambient noise), the viewing conditions, the display device calibration.

General environment

For the brightness conditions, all light sources, other than those used for room lighting (fluorescent tubes of controlled variable intensity), should be avoided as they significantly degrade video quality. The screen should be positioned in such a way that no light source, such as a lamp or window, is directly in the viewer's field of view, or may cause reflections of certain surfaces on the screen. We realized our tests in quite and acoustically processed rooms conforming with the ITU-T P.911. The noise level was below 30 dBA with no dominant peaks in spectrum.

Viewing conditions

The viewing distance has a direct influence on video perception. In fact, the distribution of the spatial frequencies of the video projected on the retina depends on this distance. As recommended in ITU-T P.910 we placed the display screen in a distance equal to $3 \times H$ (screen hight) from the subjects. The control of the ambient brightness is important because there is only a small part of the visual field that is excited by the displayed video stream, the rest is by the environment. Thus, we adapted the ambient brightness of the rooms in order to limit the glare and the visual fatigue of the observers.

Display device calibration

Calibrating a screen consists of making four important settings.

- The maximum brightness of the screen (white point).
- Gamma.
- Color temperature (in Kelvin).
- The minimum luminosity (Black point).

In order to calibrate our display devices, we have used a tool to neutralize the display defects of the screen and to automatically adjust the hardware settings (brightness, contrast, white point, etc.) so that the display device ensures that it displays the widest range of possible colors.

Figure 3.3: Display device calibration

3.1.3 Global conduct of the sessions

The overall structure of a test session is common to most of the methodologies we used. The procedure for presenting the sequences is specific to each test context and will be described in detail later. The main steps of a subjective audiovisual quality session in chronological order are:

- acuity and colors perception tests,
- instructions,
- training session,
- main test.

After the selection of observers meeting the visual testing criteria, a session is preceded by an explanation of the type of methodology, the scoring system, the presentation protocol and any useful elements. The psychological conditions in which the observer is placed are both difficult to define and very influential on his assessment, which gives great importance to this preliminary explanation and instructions. As recommended in ITU-T Rec. P.910, we started the test with

a few typical conditions to anchor the judgment of the observers. The scores of this training session are not taken into account in the final results. At the end of the session, an individual assessment is carried out in order to detect any possible misapprehensions.

A test session consists of a variable number of sequences, which corresponds to the evaluation of a perceived quality under different conditions. Our sequences have generally a duration between 8 and 10 seconds in order to leave a sufficient time for observers to give a stable score. During the test session, the sequences are presented in a random order. Indeed, when a sequence of good quality follows a sequence of poor quality, it will be over-evaluated. These context effects are limited by a random sequencing of the sequences. In total, we respected the fact that a test session does not exceed 30 to 60 minutes, including explanations and the training session.

3.2 Statistical methodology

During a subjective quality assessment, a significant amount of data is collected. It is then necessary to carry out some tests before translating this data into results. Thus, inter-observer coherence is evaluated. As a result of this verification, the assessment scores of some observers may be rejected. This step can therefore be critical in obtaining the results of a methodology since it requires a minimum number of observers. Once the inter-observer coherence of the results has been verified, synthesis tools are used to draw conclusions. Simple statistical tools are often used, but depending on the type of test, more advanced tools may be useful. Here we present algorithms for the subjects screening and statistical tools for synthesizing results.

3.2.1 Subjects screening

After collecting the subjective scores, it is essential to validate in order to eliminate all subjects whose data might be biased. Multiple reasons may be the cause of invalid subject's scores, including lack of concentration of the subject, failures on the part of the experimenter, the video playback system, or the rating save system. Thus, a screening method must be applied to remove the outliers and to only retain subjects who are able to rate video sequences consistently. In our analysis, we used the two following screening algorithms. We rejected observers that are discarded by both of the algorithms.

ITU-R BT.1788 (SAMVIQ)

ITU-R BT.1788 [128], also known as SAMVIQ, demands that subjects have a stable and coherent method to vote degradation of quality. This technique rejects subjects who do not associate with other subjects (i.e., rank impairments differently). The rejection criteria uses the linear correlation coefficient of Pearson between x and y:

$$r_p(x,y) = \frac{\left(\sum_{i=1}^{N_c} x_i.y_i\right) - \frac{\left(\sum_{i=1}^{N_c} x_i\right)\left(\sum_{i=1}^{N_c} y_i\right)}{N_c}}{\sqrt{\left(\sum_{i=1}^{N_c} x_i^2 - \frac{(\sum_{i=1}^{N_c} x_i)^2}{N_c}\right)\left(\sum_{i=1}^{N_c} y_i^2 - \frac{(\sum_{i=1}^{N_c} y_i)^2}{N_c}\right)}} \tag{3.1}$$

with i is the test condition, x_i the mean score of all observers on condition i, y_i is the score of an observer on condition i and N_c is the total number of stimulus ($number of conditions \times number of scenes$). The SAMVIQ screening algorithm also uses the Spearman rank correlation coefficient between these same x and y:

$$r_s(x,y) = 1 - \frac{6 \times \sum_{i=1}^{N_c} [r(x_i) - r(y_i)]^2}{N_c^3 - N_c} \tag{3.2}$$

with $r(x)$ is the ranking order of the element x. The rejection algorithm then evaluates the difference between the minimum of these two coefficients for the concerned observer and for all the other observers. If the minimum of an observer is above a certain threshold, then he is rejected.

Algorithm 1: ITU BT.1788 subject screening

if $[mean(r) - sdt(r)] > MCT.$ **then**
  Rejection threshold $= MCT$
else
  Rejection threshold $= [mean(r) - sdt(r)]$
if $[r(obs_i)] >$ Rejection threshold. **then**
  obs_i is not discarded
else
  obs_i is discarded

where r is the minimum of the Pearson and Spearman correlation; mean(r) is the average of the correlations of all the observers; $sdt(r)$ is the standard deviation of all observers' correlations; MCT is Maximum Correlation Threshold which is equal to 0.85 for SAMVIQ and DSCQS methods and equal to 0.7 for SS and DSIS methods.

VQEG Multimedia Phase I Test Plan

This screening algorithm is presented in the VQEG Multimedia Phase I Test Plan in Annex VI. The rejection criteria tests consistency of the raw scores using

Pearson correlation on both a per-clip basis and averaging scores across all scenes associated with one impairment (i.e., per-HRC or Hypothetical Reference Circuit). This technique rejects subject who do not associate with other subjects (e.g., rank impairments differently). The thresholds fixed to be appropriate for ACR tests are equal to 0.75 for the Pearson correlation per observer and to 0.8 for the Pearson correlation per HRC.

3.2.2 Correlations and statistical tests

Mean Opinion Score

Once the tests are performed, the results are analyzed and combined in a single note per video sequence describing its average quality. This note called Mean Opinion Score (MOS) is given by the following formula:

$$MOS(i) = \frac{1}{N_{obs}} \sum_{j=1}^{N_{obs}} Note_i(j) \tag{3.3}$$

where N_{obs} is the total number of participants and $Note_i(j)$ is the quality score affected to the sequence i by the observer j.

Confidence interval

A confidence interval is often associated with each MOS score, thus reducing the impact of possible errors. It is generally set at 95

$$[MOS(i) - e_j, MOS(i) + e_j] \tag{3.4}$$

where

$$e_j = 1.96\sigma_j \tag{3.5}$$

and

$$\sigma_j = \sqrt{\frac{1}{N_{obs} - 1} . \sum_{\omega=1}^{N_{obs}} (Note_i(k) - MOS(k))^2} \tag{3.6}$$

Statistical test

In order to analyze the impact of the different test conditions on the quality perception, we used of the Mann-Whitney U statistical test. We set the significant difference level to $\alpha = 0.05$. If we consider the two hypothesis:

$$\begin{cases} \mathcal{H}_0 : P_x = P_y \\ \mathcal{H}_1 : P_x \neq P_y \end{cases}$$

where P_x is the law distribution of the observation $X = (x_1, x_2, ..., x_{nx})$ and P_y is the law distribution of the observation $Y = (y_1, y_2, ..., y_{ny})$. The test involves the calculation of a static value called U. The rule decision is the following:

$$\begin{cases} \text{if} U \leq c & (\mathcal{H}_1)\text{is true,} \\ \text{if} U > c & (\mathcal{H}_0)\text{is true.} \end{cases}$$

where c is a critical value determined from the Mann-Whitney table.

ANOVA

ANOVA (ANalysis Of VAriance) is a statistical technique for comparing averages of more than two populations. In our subjective tests, we used the ANOVA method in order to test whether data from several groups have a common mean. This determines whether the groups are actually significantly different in the measured characteristic. We carried ANOVA as follows: one-way ANOVA, a simple special case of the linear model. The one-way ANOVA form of the model is :

$$y_{ij} = \alpha_j + \varepsilon_{ij} \tag{3.7}$$

Where y_{ij} is a matrix of observations in which each column represents a different group. α_j is a matrix whose columns are the group means. ε_{ij} is a matrix of random disturbances.

Principal Component Analysis (PCA)

It a multidimensional descriptive method to synthesize complex statistical data. It consists in projecting the data in a space of reduced size in order to highlight possible structures most relevant within the data. The most relevant attributes are presented according to their importance. PCA method calculates a set of variables, called principal components, representing a linear combination of the original variables. The principal components form an orthogonal basis for the space of the data. This method allow us analyzing the different modalities and factors that have the most impact on the global quality perception.

3.3 Test 1 : Non interactive videoconferencing test

3.3.1 Objectives

In this section, we present a non-interactive audiovisual quality assessment experiment conducted on audiovisual clips collected using a PC-based videoconferencing application connected via a local IP network. Through the analyses of the experimental results, we try to better understand the influence of network impairments (packet loss, jitter, delay) on the perceived audio and video qualities, as well as their interaction effect on the overall audiovisual quality in videoconferencing applications. Furthermore, our objective is to update the human perception acceptability

limits of audio-video synchronization for video conferencing. We investigated the contribution of this synchronization to the audiovisual quality independently and accompanied with network impairments. Finally, we propose an integration model to estimate the audiovisual quality in the studied context.

3.3.2 Related work and motivation

Taking into account the multi-modality of an audiovisual content, it is essential to consider the interaction between audio and video qualities in order to evaluate the human audiovisual quality perception. Previous studies have shown that individual audio and video qualities influence the perceived audiovisual quality but not with the same degree. Indeed, it depends on many factors such as the subject attention, the usage context, the audiovisual content or the experimental environment [129, 130]. For content corresponding to news, teleconference or music clip, the audio stream quality has greater weight on the overall quality [131]. In addition, some studies have shown that there is a significant mutual interaction between the video and the audio quality [132].

Models have been proposed for various types of contents and different types of degradation [130, 131]. However, there are few studies addressing the impact of the network settings on perceived multimedia quality [130, 133]. In [134, 135], the authors studied the quality of multimedia content and they found that both auditory and visual qualities contribute significantly to perceived multimedia quality, but they did not take network errors into consideration in their proposed models.

Another factor that considerably influences the perceived quality is the audio-video synchronization. Most of the studies that investigated this problem are old [136] and the proposed acceptability threshold must be updated to be more adapted to current solutions. Nowadays, the habits of using video communication services by customers have changed, their requirements are evolving and technologies of video restitution are advancing.

Furthermore, most of studies that proposed models for estimating multimedia quality focused on synchronized contents [129]. They analyzed the impact of network and application impairments separately from the audio-video synchronization. The impact of the packet loss on audio-visual communication was well investigated in [137, 138]. These studies concerned the synchronization problem caused by the packet loss, but only in IPTV scenarios (not for videoconferencing applications). In addition, the combination of the network impairments and non-synchronized audio and video has not been well studied in the literature.

Through this subjective test, we study the impact of the audio and video qualities on overall audiovisual quality in the context of video telephony on PC; propose new acceptability thresholds of audio-video asynchrony in video telephony context;

and study the effect of synchronization in the presence and absence of network degradation.

3.3.3 Experimental set-up and recording

In order to generate our test database, we used a video conferencing software internally developed in Orange. The reason of this choice is that it allows sharing multimedia contents between two users and separating audio and video IP flows. Thus, we were able to simulate degradations on audio and video independently. We used the audio-visual communication protocol H.323 recommended from the ITU [139] to transmit calls between two users.

To simulate network degradations, we used the NetDisturb software [140] which allows disturbing flows over IP network by generating user-defined impairments (latency, jitter, packet loss . . .). The interest of using a network simulator, instead of the real network, in our experience was to totally control the IP network degradation, provide repeatable QoS on audio and video flows using predefined configuration mode and values, and re-create real world problems in the laboratory. We inserted a machine equipped with NetDisturb between our two clients connected via an Ethernet local network.

Once a conference call was set up, the client sender transmitted the original audio, video or audiovisual files to the receiver (see Fig. 3.4). Then, we controlled the packets transmission between them by adding packet losses, jitter and delay. At the receiver side, we recorded the degraded sequences and captured IP packets traveling over the network (pcap format). To ensure a perfect playback, all recorded multimedia sequences were processed and stored as raw YUV 4:2:0 for the video stream and uncompressed Pulse Code Modulation (PCM) for the audio stream.

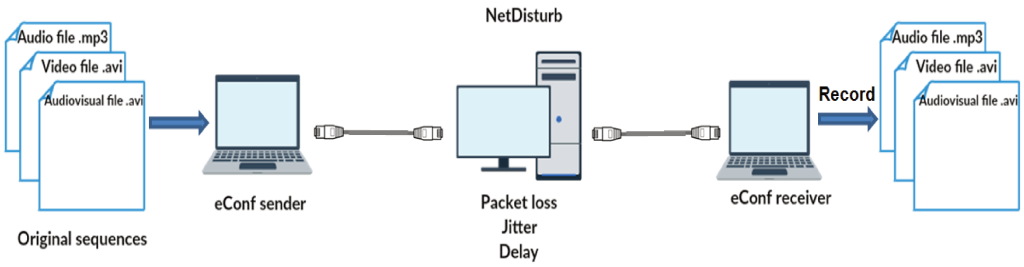


Figure 3.4: Simulation platform design.

3.3.4 Conditions

The distortions we simulated reflect the range of IP network impairments including packet loss, jitter and delay. The level of distortion was varied to generate multime-

dia contents at a broad range of quality from hardly perceptible to highly annoying levels of impairments, as recommended in [141]. Applying different network delays for audio and video streams generated the asynchrony between them. A negative value of the delay means that the audio stream is delayed according to the video and a positive value means that it is advanced.

The values of the packet loss and jitter shown in Table 3.1 below had been set empirically (by experts who observed the results on the contents and selected those which represent an actual case of use). The percentage of the packet loss is calculated on the basis of 100 received packets. Packet loss patterns differ from network to network and over time. Due to time constraints, in our study we just tested one of two major model categories: random loss. We note that the percentage of audio packet loss is more important than the video packet loss because of the stationarity of the audio codec and the low complexity of its correction mechanisms. The jitter applied on video stream is more important that the audio stream because the size of video packets is much more important than the size of the audio packets. The asynchrony values were decided based on prior knowledge [142].

Video packet loss VPL (%)	0, 0.5, 1, 2
Audio packet loss APL (%)	0, 2, 5, 20
Video jitter (ms)	0, 60
Audio jitter (ms)	0, 30
Audio-video asynchrony (ms)	-400, -250, -150, 0, +50, +150, +400

Table 3.1: Experiment parameters

3.3.5 Source sequences

For the experiment, six sequences were selected to represent different contexts of real life video calls (Restaurant, Desk, Sofa, Poster, Hall and Park)(see Fig. 3.5) filmed by Orange teams. The audiovisual sequences are characterized by different properties of audio and video contents.

- **Restaurant** scene represents a man that makes a video call in a dining hall: high complexity (a lot of details, and noisy background).
- **Desk** scene represents a woman that makes a video call in a private environment (office): low complexity (not much movement), few details and texture (solid color jacket, white wall) sound ambiance quiet.
- **Sofa** scene represents a man that makes a video call in a private environment (sofa): average complexity (few movements, few details and texture (striped wall, yellow pillow), sound ambiance quiet or little noisy).
- **Poster** scene represents a woman making a video call in a private environment (office) and showing a poster to her interlocutor: high complexity (a lot of

Figure 3.5: Frame captures from the original sequences.

movement (moving camera), a lot of details (text), sound ambiance quiet).

- **Hall** scene represents a man making a video call in a public place (hall of the company): high complexity (a lot of movement, a lot of details and texture).
- **Park** scene represents a man that makes a video call in a public garden: high complexity (a lot of movements, details and a very noisy background).

The duration of the sequences is between 8 and 10 seconds. The sequences represent different levels of spatial and temporal complexities (Fig. 3.6). The spatial perceptual Information (SI) indicates the amount of spatial detail of a picture. Greater the value of SI, more the scene is spatially complex. As described in P.911, SI is based on the Sobel filter and it present the maximum value of the standard deviation over the pixels in each Sobel-filtred frame:

$$SI = max_{time}\{std_{space}[Sobel(F_n)]\} \tag{3.8}$$

The temporal perceptual Information (TI) indicates the amount of temporal changes of a video sequence. More the sequences contains high motion higher the value of TI is. The measure of TI is computed as the maximum over time of the standard deviation over space of $M_n(i,j)$ over all pixels in the positions i and j (P.911).

$$TI = max_{time}\{std_{space}[M_n(i,j)]\} \tag{3.9}$$

where $M_n(i,j)$ is the difference between pixels at the same position in the frame,

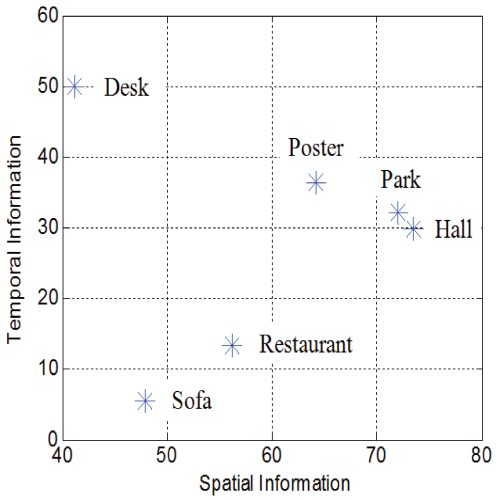


Figure 3.6: Spatial (SI) and temporal (TI) perceptual Information of the source sequences.

but belonging to two subsequent frames:

$$M_n(i,j) = F_n(i,j) - F_{n-1}(i,j) \tag{3.10}$$

where $F_n(i,j)$ is the pixel at the i^{th} row and j^{th} column of n^{th} frame in time.

The source sequences' resolution is VGA (640×480) with a frame rate of 15 fps as typically used in video conferencing mobile applications. Before simulating the network impairments, we used the FFMPEG tool [143] to encode the video stream to H.264 codec format at the bit rate of 768 kbps. This codec is used with this bit rate because the resulting encoded stream has a good quality (PSNR=60.35; SSIM=0.97). The audio stream of the sequences was coded with AMR Wide Band codec at 23.85 kbps (1 channel). This codec and bit rate value was chosen because it ensures a good perceived quality. The experimental conditions are summarized in Table 3.2.

3.3.6 Methodology and test protocol

Among the objectives of this experimental study is to investigate the impact of audio and video single stream quality on the overall audiovisual perceived quality and their mutual influence. Thus, quality scores of the separate audio and video streams must be collected. To do so, our experiment was organized in three sessions as detailed in Table 3.3. The experimental method was the Absolute Category Rating (ACR). The test protocol was based on the recommendations ITU-T P.800 [144], ITU-T P.910 [145] and ITU-T P.911 [146] for the audio-only test, the video-only test and the audiovisual test respectively. Then, we adapted these recommendations to our specific purpose and limitations.

Video		**Audio**	
Codec	H.264/AVC (constrained baseline)	Codec	AMR Wideband
Bit rate	768 kbps	Bit rate	23.85 kbps
Resolution	VGA (640 × 480)	Channels	1
Frame rate	15 fps	Sampling frequency	48000 Hz
GOP size	10 frames		
Video color scheme	16 bit YUV (4:2:0)		

Table 3.2: Experimental conditions used in the subjective study

In the audio-only test, after each presentation the subjects were asked to evaluate the audio quality(MOS_A) whereas in video-only test they were asked to evaluate the perceived video quality (MOS_V). In order to study the impact of individual audio and video qualities on the overall quality, in the audiovisual test, subjects assess audio and video qualities (MOS_A^{AV} and MOS_V^{AV}) beside overall audiovisual quality (MOS_{AV}). Subjects also evaluated the audio-video asynchrony (MOS_{synch}) in the audiovisual test.

To measure the perceived quality, a subjective scaling method is required. For the video, audio and audiovisual quality, we used a five-level MOS scale and for the synchronization, we used a specific 5 point impairment scale (see 2.3.1).

Test	Duration	Sequences	Conditions	Outputs
Audio only	10min	36	5	MOS_A
Video only	10min	36	5	MOS_V
Audiovisual	1h30	176	33	MOS_{AV} MOS_A^{AV} MOS_V^{AV} MOS_{synch}

Table 3.3: Test organization

A total of 30 subjects (13 male, 17 female) participated in the experiment. We realized the audio-only and the video-only test in the same session with 15 subjects while the audiovisual test was carried out with the other 15 subjects. They were provided with a high quality headphone (Stax SR-404) for sound reproduction. The experiment was performed in an acoustically treated room especially designed for audio and video quality tests. The signals were presented to the subjects via an LCD computer monitor with a 1024 × 768 resolution. The evaluation score was indicated on a tablet next to the screen on the right of the subjects.

Subjects were carefully introduced to the assessment method, the impairment types, the opinion scale, the stimulus presentation and timing before the start of

the experiment. The test session was preceded by a training session lasting 5 minutes. The range and type of impairments were presented in training session, which contained some sequences from those used in the test session. In the audio-only and audiovisual tests, subjects were allowed to adjust the playout to a comfortable level from the sound card during the training session, but not during the test session. Subjects were allowed to take breaks when they feel tired.

3.3.7 Results analysis

The test results were summarized by computing the averaged MOS values for each test condition over the six sequences and the confidence interval (CI) of the estimated mean. Before calculating the MOS subjective scores, we processed to the screening of the subjects. We used the algorithm described in [147] in order to detect and eliminate possible outliers. Our screening results show that no subject has to be excluded.

We used of the Mann-Whitney U statistical test in order to analyze the impact of the different test conditions on the quality perception and the interaction between the audio and video streams and their impact on the overall quality perception. We used this statistical test because our data does not follow the normal distribution.

Audio-video quality Interaction

The plots in Fig. 3.7 show the MOS scores averaged over all sequences for both test sessions. They demonstrate that the experiments have been properly designed, as the subjective rates uniformly span over the entire range of quality levels. By plotting MOS_V vs. MOS_V^{AV} and MOS_A vs. MOS_A^{AV}, and calculating their linear correlation coefficients ρ, we noticed that the perceived audio and video qualities are weakly influenced by the audiovisual context. Performing Mann-Whitney test on the audio and video MOS revealed that there is not a significant difference between scores of the two sessions (audio-only and video-only vs. audiovisual) (MOS_A vs. MOS_A^{AV}, p_{value}= 0.936; MOS_V vs. MOS_V^{AV}, p_{value}= 0.924). Thus, subjects rate the quality of the audio and video streams when they are separated in the same way as when they are coupled.

We are also interested in studying the mutual interaction between the individual audio and video streams. A statistical test revealed that in an audiovisual context the impact of the video impairments on the perceived audio quality is not significant (p_value= 0.665). On another hand, the audio impairments have a small impact on the perceived video quality. For the same video quality level, MOS_V^{AV} values decrease slightly with the percentage of audio packet loss. This drop in MOS scores is more significant in the case of good and average video quality levels (0%VPL, 0.5%VPL). When the video quality is already poor (1%VPL and 2%VPL), quality

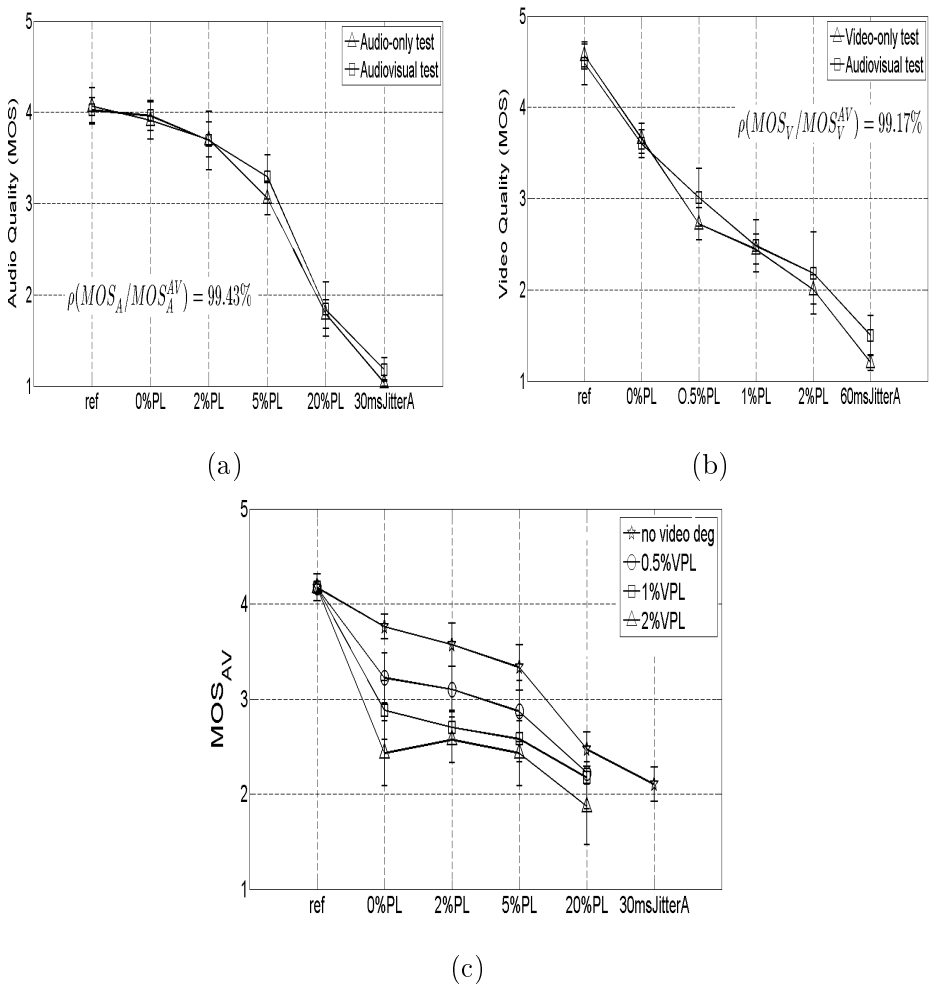


Figure 3.7: Mutual interaction between audio (a) and video (b) qualities and the impact of audio and video quality on overall audiovisual quality (c).

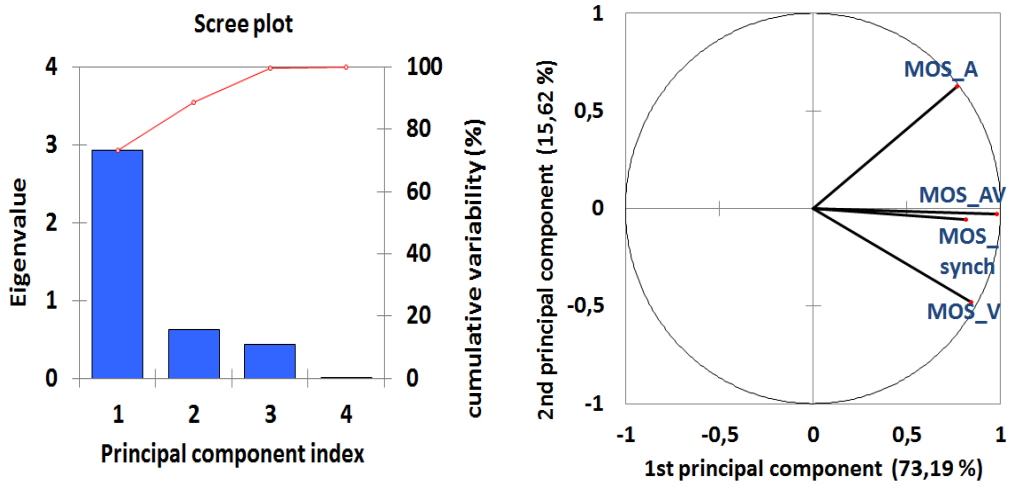
judgment is not affected by the audio degradation (there is not a significant difference). For a non-interactive evaluation of video telephony content, a similar study also revealed a small but not significant mutual influence between audio and video qualities [148].

Fig. 3.7 (c) shows the interaction between audio and video quality levels in influencing the overall audiovisual quality. The presented results were averaged over all delays (synchronous and not synchronous contents) and over all contents. An analysis of the subjective data reveals that for the same audio quality levels, decreasing the video quality generally results in inferior audiovisual ratings. Alongside, for the same video quality, decreasing the audio quality generally results in inferior audiovisual ratings. The impact of video impairments on audiovisual quality at good audio quality level is more significant than at poor and bad audio quality levels. Con-

cerning the jitter condition, it had the biggest impact on decreasing the perceived quality. Sequences generated with jitter impairment presented the lower audiovisual quality. Thus, we measured the impact of jitter independently, without crossing it with packet loss in order to not bias the subjective results.

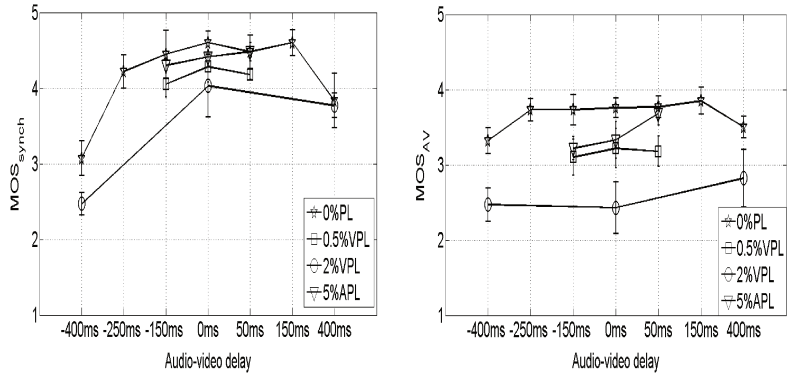
In order to study the influence of audio quality AQ, video quality VQ and the synchronization on the overall audiovisual quality AVQ, we performed a Principal Component Analysis (PCA). We constructed four dimensional test vector composed of MOS_A^{AV}, MOS_V^{AV}, MOS_{synch} and MOS_{AV}. In Fig. 3.8(a), we represent the eigenvalues corresponding to the four principal components. The first two components account for 88.81% of the variance. The PCA results from Fig. 3.8(b) show the influence of individual modalities on the overall audiovisual quality. From these results we are able to conclude that both AQ and VQ contribute to AVQ. It can be observed that the synchronization is an important factor that impacts considerably the perception of audiovisual quality. Thus, it is essential to more investigate the relation between the synchronization values and the quality evaluation.

We note that generally the video quality influences the overall audiovisual quality more than the audio quality as revealed in [146]. The Pearson correlation between MOS_V^{AV} and MOS_{AV} is equal to 87.6% while the correlation between MOS_A^{AV} and MOS_{AV} is equal to 75.6%. On another hand, the Pearson correlation between MOS_{synch} and MOS_{AV} is equal to 70.1%.



(a) (b)

Figure 3.8: Principal Component Analysis



(a) (b)

Figure 3.9: Synchronization acceptability chart.

Audio-video Synchronization

From the test results we were able to identify thresholds of synchronization acceptability. We set MOS_{synch} score equal to 4 and MOS_{AV} equal to 3 as the acceptability limits, corresponding to thresholds when the subjects begin to be disturbed and when the audiovisual quality becomes poor. For audio delayed with more than 250 ms and advanced with more than 150 ms, the desychronisation becomes annoying and the audiovisual quality decreases. These results are consistent with limits reported for IPTV (i.e. on large TV screens) [137, 138, 142], and show thus that the screen resolution has almost no influence on the perception of desychronisation between audio and video. Limited by the number of conditions and the duration of the test, we could not cross the 6 different values of delay with all the network degradation levels which explained the lack of some points on the graphs. Furthermore, as shown on Figure 3.9, the presence of video (and in a smaller extent of audio) packet loss impairments have a little, but not significant impact on synchronization. The perception of audiovisual quality and synchronization is sensitive to network degradation mainly related to video streams.

Audiovisual quality model

Stepwise linear regression models were applied to study the influence of audio and video qualities on audiovisual quality. The general model proposed in conventional studies [134, 149] was assumed as follows:

$$MOS_{AV} = \alpha_0 + \alpha_1 MOS_A^{AV} + \alpha_2 MOS_V^{AV} + \alpha_3 MOS_A^{AV} \cdot MOS_V^{AV} \quad (3.11)$$

In our study we propose a new prediction model by integrating the desynchronization term $DMOS_synch = 5 - MOS_{synch}$ since the PCA results show that the synchronization has an important impact on the perception of the audiovisual quality.

$$MOS_{AV} = \alpha_0 + \alpha_1 MOS_A^{AV} + \alpha_2 MOS_V^{AV} + \alpha_3 MOS_A^{AV} \cdot MOS_V^{AV} + \alpha_4 DMOS_{synch} \quad (3.12)$$

The correlation results of regression analysis are summarized in Table 3.4.

Model	R^2
MOS_A	0.57
MOS_V	0.77
$MOS_A + MOS_V$	0.94
$MOS_A \cdot MOS_V$	0.95
$MOS_A + MOS_V + MOS_A \cdot MOS_V$	0.94
$MOS_A + MOS_V + DMOS_{synch}$	0.94
$MOS_A \cdot MOS_V + DMOS_{synch}$	0.96
$MOS_A + MOS_V + MOS_A \cdot MOS_V + DMOS_{synch}$	0.94

Table 3.4: Linear correlation of models

From the shown results, we note that the multiplicative with the synchronization term model solely provides the best fit. We applied this model to the subjective test results and we found a linear correlation (R^2) between subjective and estimated qualities equal to 96.6% and a root mean square error ($RMSE$) equal to 0.13. The mean of the 95% confidence interval (MCI) for the subjective MOS was 0.22. The evaluation error of the model was less than the statistical ambiguity of the subjective score (i.e., $RMSE < MCI$), so the quality evaluation accuracy of the model is sufficient for practical use. By applying multiple regression analysis we determined the constants. Thus, our model is the following:

$$MOS_{AV} = 1.57 + 0.16 MOS_A^{AV} \cdot MOS_V^{AV} - 0.15 DMOS_{synch} \quad (3.13)$$

In comparison, early studies already suggested multiplicative models [146, 150] but all of them were based on synchronized contents, while the model we proposed here takes into account the asynchrony and is based on pure network impairments.

3.3.8 Summary

In this section, we presented an audiovisual quality study for videoconferencing in the presence of IP network transmission errors and extended the scope by introducing asynchronous contents. The results showed that both audio and video quality contribute to the overall audiovisual quality with a general domination of video quality. We proposed an integration model to predict multimedia quality which takes into account desynchronized contents with pure network impairments. In the

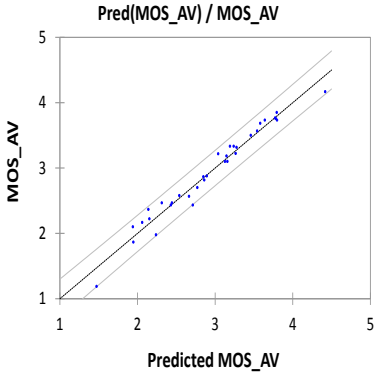


Figure 3.10: Predicted vs. MOS_{AV} model from Eq. 1.3 with 95% confidence interval .

next section we study further the questions of audiovisual quality and asynchrony in a more realistic interactive context.

3.4 Test 2 : Interactive videoconferencing test

3.4.1 Objectives

Through this subjective test we try to discuss the following questions:

Q1: What is the impact of scene complexity on the perceived audio, video and audiovisual qualities?

Q2: What is the impact of scene complexity on audio-video desynchronization acceptability?

Q3: Are the perception of audiovisual, audio and video qualities the same in non-interactive and interactive contexts?

In our experiment, we are also interested in audio and video synchronization since it is a factor that considerably influences the perceived quality of multimedia services [151]. Thus, we precise and compare the thresholds of asynchronization acceptability between the non-interactive and the interactive contexts.

3.4.2 Related work and motivation

Interactive conversational subjective experiments are closer to a real-life video-telephony, or video-conferencing calls than non-interactive experiments. However, current audiovisual assessment researches mainly focus on non-interactive applications, such as video-on-demand, streaming or IPTV services [135, 130, 131]. Few studies have been conducted for evaluating audiovisual quality in conversational context [152, 150, 153].

In [154], the authors assessed the perceived VVoIP (Voice and Video-over-IP) conversational quality under different network conditions (packet loss and delay). On the other hand, some recent studies were interested in determining the factors that impact the perceived conversational quality. For example, in [152] the authors studied the influence of the conversational scenario and communication task on the perceived quality. They showed that the subjects' concentration on audio or video quality depends on the type of scenario, which influences their judgment. In [150], the influence of the experiment context on the audiovisual modalities was also investigated. Through their comparative study, the authors concluded that for test conditions of low audiovisual quality level, the MOS scores collected in an interactive context are greater than the ones collected in a non-interactive one. On the opposite, when the level of audiovisual quality is high, the experimental context does not seem to have an impact on the perceived quality.

All these studies focused on the impact of the conversational task and the scenario on the subjective scores. However, they do not explore the effect of the scene complexity on the perceived conversational quality. In real life video-conference communication, the environment around the persons varies according to their position (desk, open space, home, etc.). The differences between the environments correspond to the variation of spatial and temporal complexity of the scene. Our contribution consists in studying the influence of the scene complexity and the test context on the perceived audiovisual quality under certain network transmission conditions.

3.4.3 Experimental set-up and recording

To perform this test under a controllable environment, we used an internal video conferencing software. Figure 3.11 depicts the videoconferencing test bed configuration used for the experiment. User PC1 and user PC2 are two identical videoconferencing systems (hardware and software), running our videoconferencing software, placed in two separate rooms and connected via a local Ethernet IP network. They were used by the subjects to make video calls. The audio-visual communication protocol H.323 [139] was used to transmit calls between two users.

Both terminals were controlled remotely. Network degradations (packet loss and delay) introduced on the video or audio streams were realized from a remote room. To introduce different sets of packet loss for audio and video streams and to generate audio and video delay, we inserted in the transmission path a machine equipped with the network simulator "Netdisturb".

The video conversation window was shown in the VGA (640×480) resolution. The video and audio setting (codecs and bit rates) are the same as the non interactive test and were unchanged throughout the test. During every conversation, we took simultaneously a screen record of the multimedia communication contents. We also

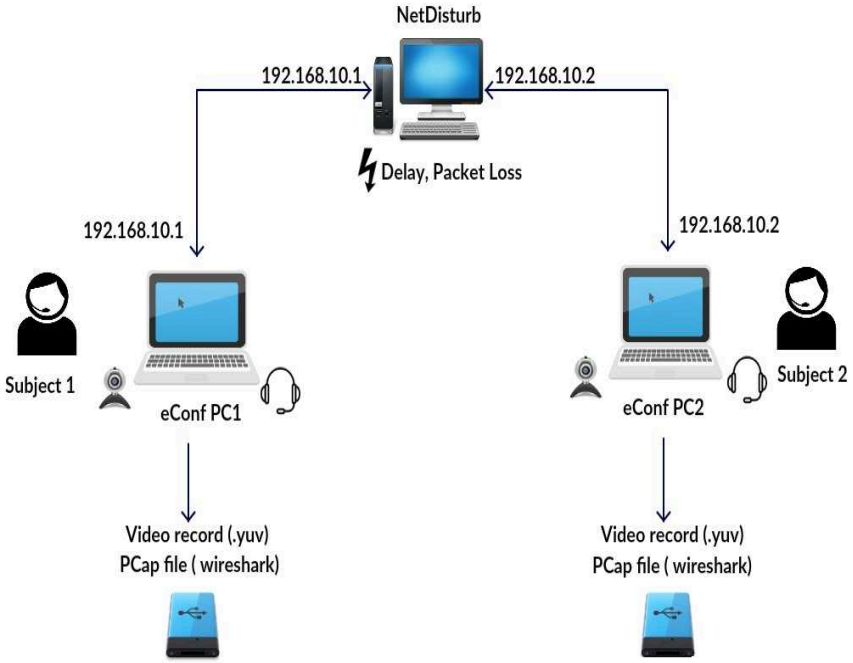


Figure 3.11: Simulation platform design.

capture IP packets transmitted over the network (PCAP files using Wireshark). To ensure a perfect playback, all recorded multimedia sequences were processed and stored as raw YUV 4:2:0 for the video stream and uncompressed Pulse Code Modulation (PCM) for the audio stream.

3.4.4 Conditions

We simulated IP network impairments including packet loss and delay. We generated two levels of audio and video packet loss which represent the extreme ranges of quality: 1- hardly perceptible, 2-highly annoying. The configured distribution law of packet loss rates was random. All conditions were symmetric so that the test participants experienced the same quality on both ends of the connection. We randomized the order of the conditions. Table 3.5 provides an overview of the transmission parameters evaluated in this study.

3.4.5 Methodology and test protocol

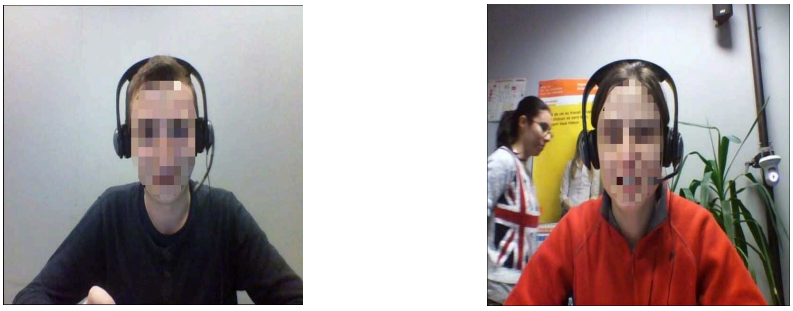
In order to consider the influence of scene complexity and keep the experiment time within limits, we have only configured the two rooms with two different levels of video complexity:

Video packet loss VPL (%)	0, 0.5, 2
Audio packet loss APL (%)	0, 5, 20
Audio Delay AD (ms)	0, 250, 400
Video Delay VD (ms)	0, 150, 400

Table 3.5: Experiment conditions

- **Room 1**: where the background behind the subject is a simple white wall (see Fig. 3.12.a).
- **Room 2**: where the scene has a certain spatial and temporal complexity. A poster and a plant behind the subject, and one Orange staff walk behind him from time to time (see Fig 3.12.b).

The rooms have been acoustically treated and they have a similar audio background.



(a) (b)

Figure 3.12: Screen captures of the conversation in Room 1 (a) and Room 2 (b).

The test has been conducted in an interactive scenario. We proposed a game to stimulate the conversation between the two subjects. For a subject, the objective of the game, was to let its partner guess a word without using the word itself or five additional words listed on a card. We gave each subject 20 cards. This conversation task is similar to the Name-Guessing task from the ITU-T Recommendation P.920 [60]. The subjects could also discuss on their own topic if they prefer. The duration of each conversation was around three or four minutes. Each discussion corresponds to a specific set of impairments of audio and video. The subjects tested 9 different conditions where the audio and video impairments are independent (limited by the test duration, the interaction between the conditions was not tested).

Twenty subjects (9 male, 11 female) participated in the experiment. They were all inexperienced in evaluating audiovisual quality in such a context, but the majority had already experienced a video-conference call. Each subject was individually

briefed about the goal and the procedure of the experiment. A training session of 3 minutes preceded the actual test. The purpose of this session was to make the subjects familiar with the testing procedure and the variations of audio and video quality. During the training the IP flow was impaired by the same type of distortions as the main test.

In this experiment, subjects were asked to rate the perceived overall audiovisual quality (MOS_{AV}), audio quality (MOS_A) and video quality (MOS_V) as well as the audio-video synchrony annoyance (MOS_{synch}). An absolute category rating (ACR) was used for collecting subjective quality judgments. The subjects rated the qualities and the synchronization using the five-grade scales presented in 2.3.

3.4.6 Results analysis

The results of the subjective experiment are summarized by averaging the scores assigned by the panel of participants for each conversation. We calculate the Mean Opinion Score (MOS) and the corresponding Confidence Interval (CI).

For the comparison between the experiment interactive and non-interactive contexts and between the scene complexity, the Mann-Whitney U test [155] is used since the data does not follow the normal distribution. We set the significant difference level to $\alpha = 0.05$.

Prior to the MOS computation, a screening of the subjects is preceded using the algorithm described in [156] in order to detect and exclude possible outliers, that is, subjects whose evaluation significantly deviates from others. Our screening results show that no subject has to be excluded.

Influence of the experiment context: interactive vs non-interactive

In this section, we investigate the influence of the experimental context (non-interactive vs. interactive) on audiovisual quality (AVQ), video quality (VQ), audio quality (AQ) and audio-video synchronization acceptability. Figure 3.13 shows the MOS scores obtained in the two contexts averaged over all scenes. By comparing the two plots in Figure 3.13.a, we observe that there is a significant difference between MOS_{AV} scores in case of 0.5% video packet loss and 20% audio packet loss. This may indicate that subjects are more sensitive to low video impairments when they communicate than when they passively watch an audiovisual sequence. The interactive task may make the subjects discriminant and severe in the assessment of the audiovisual quality since the video impairments may have more psychological impact on the visual communication they are involved in. However, for important VPL (2%) the quality is poor enough that there is not a significant difference between subjective scores in the two contexts. The subjects give a significantly higher quality note in the interactive context than in the non-interactive context when the

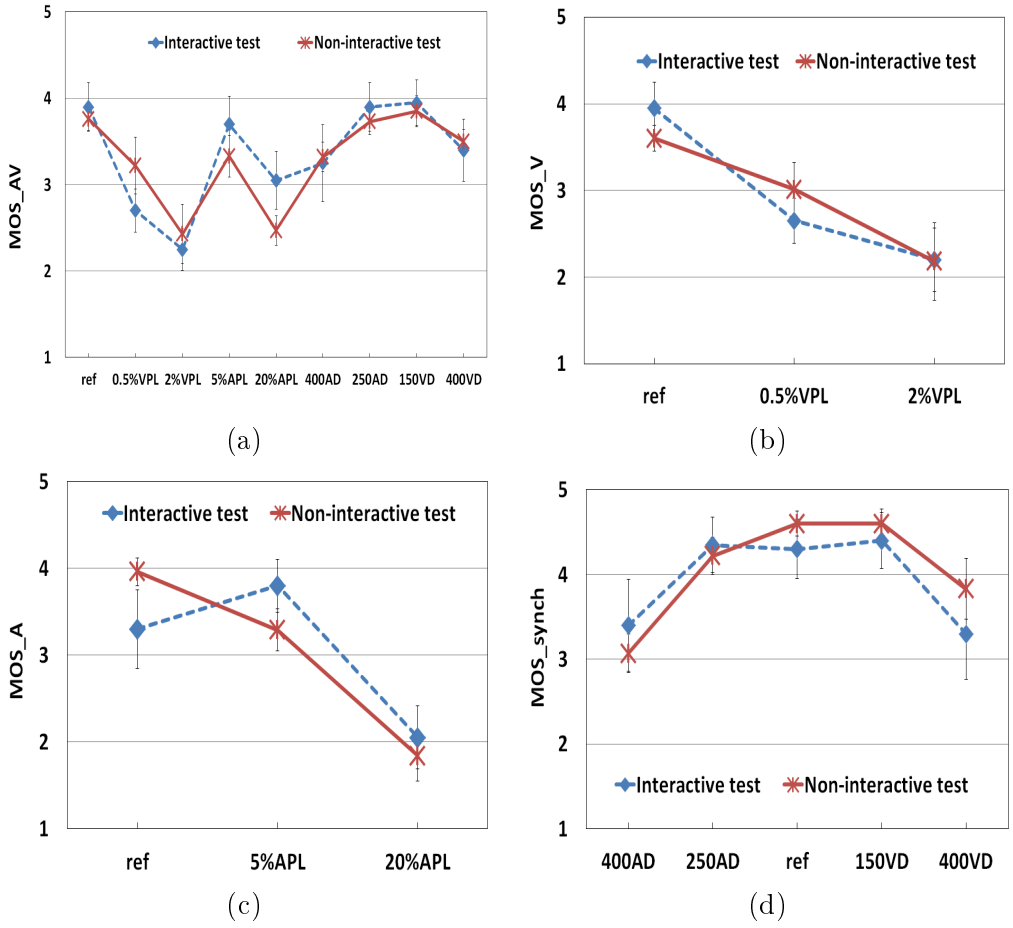


Figure 3.13: Interactive vs. non-interactive MOS scores

audio quality is very low. This may indicate that their attention on the audiovisual quality judgment may be diverted by the guessing game.

For the perceived video quality, Figure 3.13.b shows no significant difference between the two contexts. Subjects perception of the video quality and concentration on the artifacts are the same. Nevertheless, we report a significant difference of perceived audio quality between the two contexts (Figure 3.13.c). Considering the variances, we note that for the interactive test there was not a significant difference between reference and 5%APL condition, while for the non-interactive there was this significant difference – indicating that the impairments are more noticeable in the non-interactive context. The reason of this variance may be that the audio impairments are more noticeable when the subjects are just viewing and listening to an audiovisual content. Then, they are more concentrated and they are more able to notice the impairments.

Concerning the thresholds of desynchronization acceptability, as it can be seen in Figure 3.13.d, there is not a significant difference between the two test contexts.

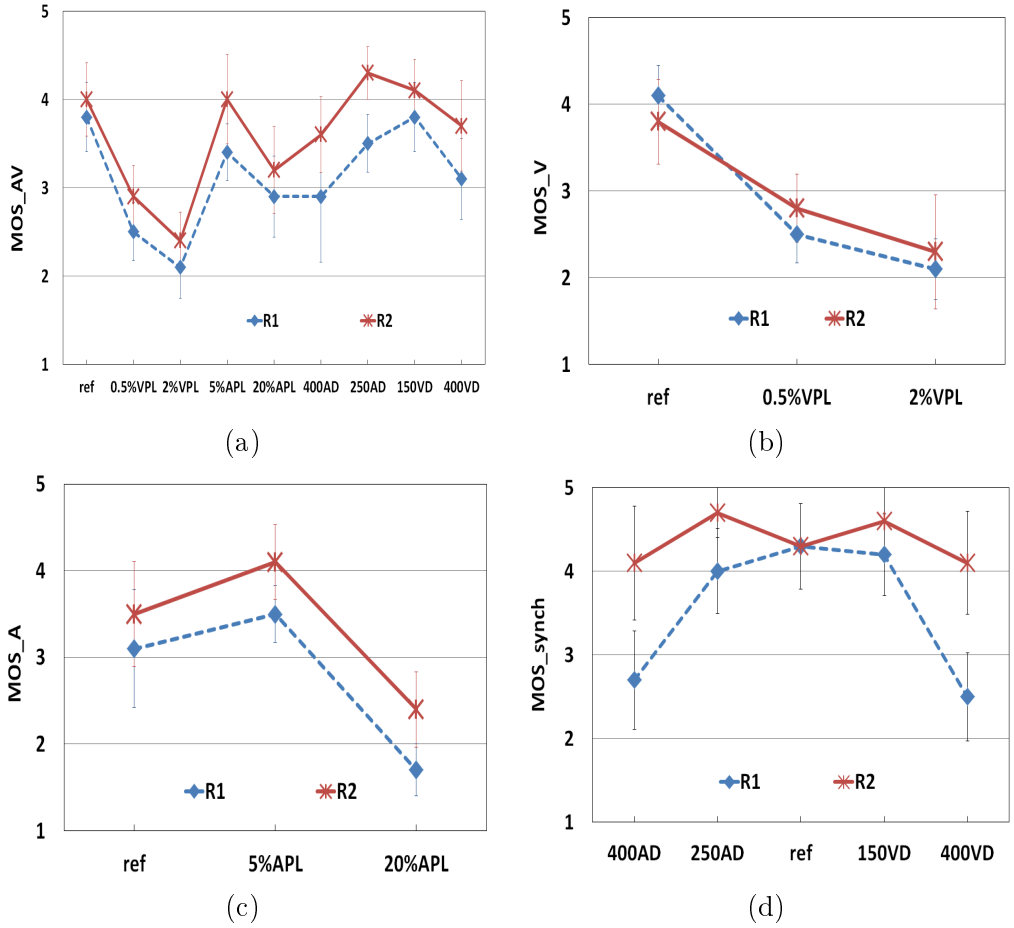


Figure 3.14: Impact of scene complexity for interactive experiment context.

These results may be true for our tested conditions where only one modality (video or audio) was impaired at the time (no interaction between the conditions). Previous studies showed that when both audio and video were impaired, differences in MOS ratings were found [150].

Influence of scene complexity for interactive experiment

In this session, we investigate the influence of the scene complexity on multimedia quality and audio-video synchronization acceptability for each experiment context.

Figure 3.14 shows the MOS_{AV}, MOS_V, MOS_A and MOS_{synch} scores associated to 95% confidence intervals, according to the quality condition and scene complexity. "R1" denotes the perception of the complex scene of Room 2 from Room 1; and "R2" denotes the perception of the simple scene of Room 1 from Room 2.

We can see that generally the perceived AVQ is higher in a simple scene than that in a complex scene at the same degradation levels (an average drop of MOS_{AV} score is about 0.5). The statistical test reveals that there is a significant difference between

subjective MOS_{AV} scores for the two rooms. This may indicate that when a scene is composed of complex spatial and temporal elements (presence of high frequencies in the picture and high amount of temporal activity), the network impairments would have a greater impact, and the artefacts (block loss and blockiness) would be more visible. In fact, scenes with high temporal and spatial complexities require more bit rate to be encoded. At a constant bandwidth, more encoding artefacts will occur and the efficiency of the packet loss concealment algorithm is reduced [157].

For the video quality (Figure 3.14.b), there is not a significant difference between Room1 and Room2. We have expected to have a significant difference in the results because the complexity of the scenes is guessed to have a stronger impact on video quality than on audio quality. This may be explained by the fact that the difference of complexity between the scenes is not sufficient to have an impact on the perceived video quality. To add a precision detail and explain this observation, we take two indicative sequences from the recorded conversations and we calculate the SI and TI indexes:

- For the complex scene : TI= 47, SI= 79
- For the simple scene : TI= 29, SI= 61

Thus, from this observation we might open a question to discuss in a future study: from which difference of scene complexity we could detect a significant difference in perceived video quality?

For the perceived audio quality (Figure 3.14.c), there is no significant difference between the results for the two rooms. This is logic since the spatial complexity is not expected to have an effect on audio quality. Furthermore, the audio background deployed in our experiment was the same when it comes to the both rooms used. Thus, the used audio background did not allow to reveal any impact in this case.

Figure 3.14.d shows that the synchronization annoyance of the subjects is also influenced by the spatial and temporal complexity of the perceived scene. The differences in MOS_{synch} are statistically significant. These plots of synchronization acceptability are coherent with the MOS_{AV} results even if the difference of MOS_{synch} between the two rooms is more important. This may indicate that an increased temporal activity has a direct impact on perceived lip synchronization since the movements disturb subject concentration.

Influence of scene complexity for non-interactive experiment

In order to stay coherent with the conversational test we present in this part a comparison between subjective results of the "Sofa" and the "Hall" scene. We chose these sequence scenes due to the difference of spatial and temporal complexity between them (see Figure 3.6) and to the similarity they have with the interactive

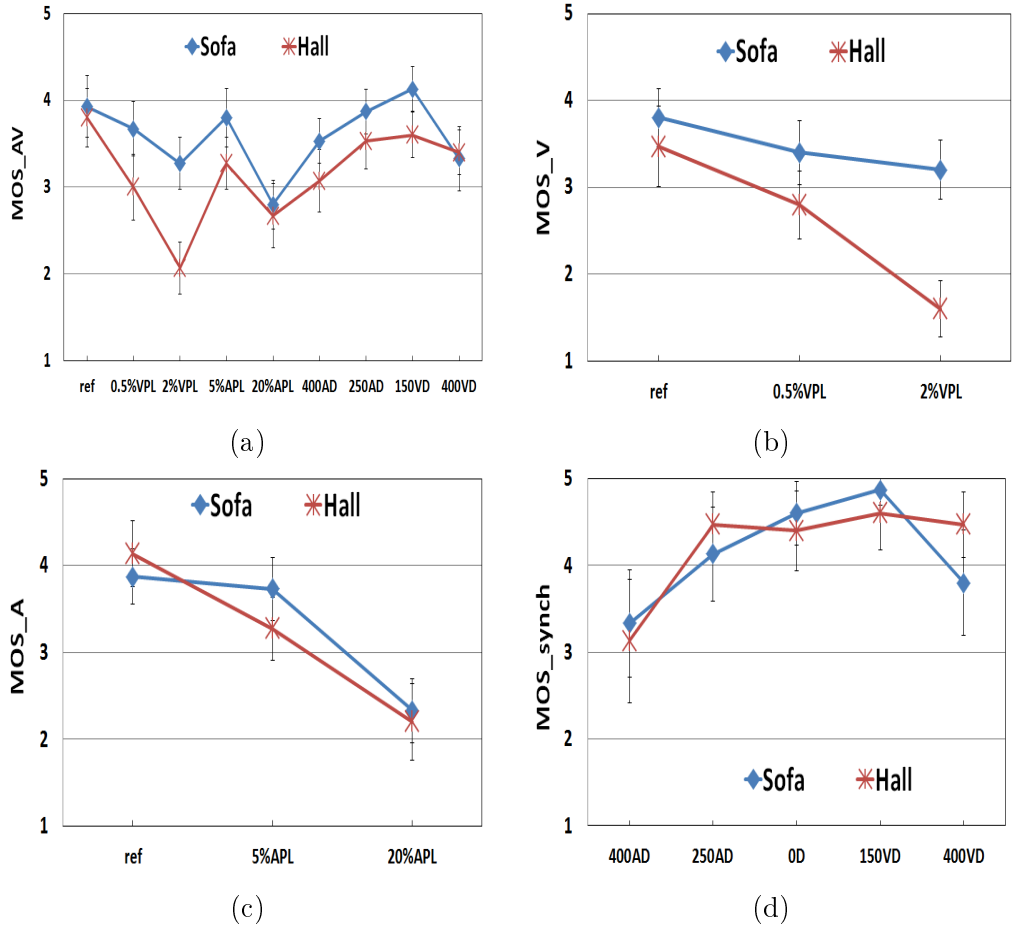


Figure 3.15: Impact of scene complexity for non-interactive experiment context.

scene content. "Sofa" sequence represents a simple scene where a guy is sitting at the sofa and talking, with white wall in its back. "Hall" sequence represents a person talking and showing a landscape (spatial complexity) in traveling mode (spatial complexity).

In Figure 3.15, MOS_{AV}, MOS_V, MOS_A and MOS_{synch} scores are represented and associated with 95% confidence intervals, according to the quality condition for each scene. We report a significant difference in MOS_{AV} scores between the sequences for all the test conditions except the reference, 20%APL and 400 ms video delay. Thus, compared with Figure 3.15.a, we deduce that overall quality perception is influenced by the complexity of the perceived scene in both interactive and non-interactive context. This observation affirms that the environment and the position of the person on the video call is a parameter to take into account to evaluate the perceived communication quality. This complexity impact could be studied through a non-interactive experiment.

For the video quality, there is a significant difference in MOS_V (Figure 3.15.b). In fact, the subjective scores of the complex scene ("Hall") are lower than that of the simple scene. This observation is justified by the fact that video artifacts caused by packet losses are more visible with sequence complexity. We notice that the SI difference between the two scenes here is much bigger than that in the interactive context. This may explain why we did not observe a significant difference in MOS_V in the interactive context.

As it can be seen in Figure 3.15.c there is not a significant difference of audio score between the two scenes. This result is expected since scene complexity has not an effect on audio quality perception, and consistent with the finding in the interactive context.

Figure 3.15.d shows that the subjects' reaction to desynchronization annoyance is the same for the two scenes, no significant difference is noticed. Thus, unlike the interactive context, in a non-interactive context the scene complexity does not impact audio-video synchronization perception. Previous studies have shown that in a passive context, large delay in the audiovisual signals does not necessarily impact the quality perception as test subjects accommodate for it [158].

3.5 Other test databases

In our subjective tests that we have described in the previous sections we have mainly studied network-type impact factors (jitter, packet loss and delay). However, there are also the application factors that impact the perceived quality of a video conference call (see Section 2.2.2) and that have not been studied. In order to complete our knowledge and to broaden the spectrum of conditions, degradation

and contents, we need to collect other bases of sequences.

In the literature most of the available and easy to access databases are for video quality tests. We choose subjective databases with variety of the included impairment types: transmission error (packet loss, jitter, freezing, etc.), coding (variable bit rates), frame rate, different error concealment algorithms. These databases will come complete the sequences of our subjective tests and will be used in two major axes of our researches:

- Evaluate the performance of the objective video metrics
- Constitute a training database of a machine learning algorithm

The characteristics of all the databases described below are summarized in Table 3.6.

3.5.1 LIVE Mobile video quality assessment Database

The Live Mobile database is developed by the Laboratory for Image and Video Engineering at the University of Texas. It's one of the most popular public VQA databases used by researchers to evaluate objective video quality assessment algorithms for wireless video transmission with regards to their efficacy in predicting visual quality. The importance of the LIVE Mobile VQA database is that it contains temporal distortions in addition to compression and packet loss distortion. In total, the distortion conditions consist of 4 conditions for H.264 compression impairments, 4 wireless-packet losses, 4 duration of frame freezes, 3 rates adapted and 5 temporal dynamics per reference. Details on these distortions are explained by authors in [159]. The videos were viewed on a mobile terminal : Motorola Atrix. The test methodology used in assessing the sequences is the single stimulus continuous quality evaluation (SSCQE) with hidden reference.

- Compression impairments: encode source videos with H.264 Scalable Video Codec (SVC) at four bit rates ($R1 < R2 < R3 < R4$) between $0.7Mbps$ and $6Mbps$. 40 distorted videos are in this category.
- Frame freezes on stored video delivery and real time live video delivery: four conditions were simulated for each source video which leads to a total of 40 distorted videos.
- Rate adaptation: change the coding bit rate during the video. We have 30 rate adapted distorted videos.
- Temporal dynamics: simulate multiple switches of the coding rate yielding 50 distorted videos.
- Wireless channel packet loss. 40 distorted videos are generated.

3.5.2 EPFL-PoliMI video quality assessment Database

The EPFL-PoliMI (Ecole Polytechnique Fédérale de Lausanne and Politecnico di Milano) video quality assessment database is freely available for download on [160]. It was specifically designed for the evaluation of transmission over IP network impairments. Packet loss distortions with different percentages (0.1%, 0.4%, 1%, 3%, 5%, 10%) are simulated in this video quality assessment test database. All sequences have been encoded with the H.264/AVC encoder adopting the High Profile.

3.5.3 SD ROI database

This database is developped by Boulos et al. in [161]. It contains videos of 6 different source contents with for each content, 14 H.264 coding conditions with or without error transmission simulations. The specificity of this database is that the spatial position of the transmission errors depends on the Region of Interest (RoI) in the video frames. The RoI are defined using an eyetracker algorithm. Then, some slice losses are introduced in the RoI and outside of it to test the impact of both the error propagation and the spatial location of the loss on the perceived quality. When the losses were outside the RoI, they occurred in the slices adjacent to the RoI. All losses were in a single I-picture to allow a longer temporal propagation.

3.5.4 SVC4QoE Replace Slice database

This database is developed by Y. Pitrey et al. in [162, 163]. It is designed for the evaluation of mobile transmission quality. It contains 9 contents with for each content, the reference (without processing or degradation) and 14 different impairment conditions. The sequences are coded with h264 and h264/SVC codecs with simulated transmission errors. Two error concealment algorithms were tested using the h264/SVC capability:

- Frame level concealment.
- pixel level concealment.

3.5.5 SVC4QoE Temporal Switch database

Developed by Y. Pitrey et al. [164, 165] this database is designed for evaluating the impact of network behavior and encoder configuration on the visual quality using SVC-based error concealment. It contains h264 and h264/SVC encoded sequences at different QP values. Several switching conditions were created between the QP values in order to test the impact of temporal quality switching on the perceived quality.

	Live Mobile	**EPFL**	**SD RoI**	**SVC4QoE Replace Slice**	**SVC4QoE Temporal Switch**
Year	2012	2010	2009	2011	2011
Nbr. of sequences	170	78	84	140	390
Nbr. of references	8	6	6	9	11
Resolution	HD 1280×720	CIF	SD (720×576	VGA	VGA
Duration	10 s	8 to 10 s	10 s	10 s	10 s
Frame rate	30 fps	30 fps	20 fps	30 fps	30 fps
Distortion types	H.264 encoding wireless packet loss frame freezes rate adaptation temporal dynamic	packet loss	Packet loss	H.264 encoding H.264/SVC encoding transmission errors	H.264 encoding
Encoder	H.264 AVC	H.264 AVC	H.264/AVC	H.264	H.264
Assessment method	SSCQE-HR	SS	ACR-HR	ACR-HR	ACR-HR
Subjective scores	DMOS $[0, 5]$	MOS $[0, 5]$	MOS$[1, 5]$	MOS$[1, 5]$	MOS$[1, 5]$
Nbr. of subjects	36	40	25	29	28

Table 3.6: Properties of subjective VQA databases

3.6 Summary

We present two modalities of subjective audiovisual quality test. This work focus on investigating audiovisual quality in interactive and non-interactive contexts and under different scene complexities. By comparing non-interactive vs. interactive test results, we summarize that statistically there is not a significant difference of MOS_A, MOS_V and MOS_{synch} scores between the two experimental contexts. Thus, in future experiments we can rely on non-interactive test results and apply them on a conversational context. However, considering MOS_{AV} scores we note a significant difference between the two contexts.

Besides, the results show that the scene complexity has an impact on the perceived audiovisual quality in both contexts and on the perception of audio-video synchronization in the interactive context. The different observation on the impact of the scene complexity on the video quality in the two contexts requires a further study. Limited by the experiment duration we studied only two different scene in the interactive context. We had not covered a wide range of spatial and temporal complexity.

Chapter 4

Perception of asynchrony: two subjective test studies

Contents

Introduction

The reduction of the temporal alignment between the auditory and visual information can alter the audiovisual perception as a multimodal event. In real time audiovisual conversation, the presence of desynchronization between the image and the sound can have a detrimental effect on the interactivity of the conversation and thus on the perceived quality. Consequently, it is necessary to control the temporal relationship between the audio and video signals so that the quality perceived by the user is not altered. As we mentioned in Chapter 2, audio/video desynchronization

is one of the most important factors to consider as the main cause of audiovisual degradation.

We have investigated the impact of desynchronization on the audiovisual quality perception in the context of videoconferencing contents in Chapter 3. The results showed that there is the same dissymmetry aspect for both TV and videotelephony applications, but with larger acceptability thresholds, rather at least 150 and 250 ms respectively [166]. The reason for this difference is not necessarily linked to the context; the design of the respective subjective tests, and in particular the question asked, can have also an impact. This is why we planned a new subjective test that included conditions with these values of delay as well as higher values, in order to see if the actual acceptability thresholds could be even higher.

Furthermore, we are interested in the interaction between different types of impairments and the audio/video synchronization that can lead to visual masking effects. In particular, we want to study if changing from high to low resolution or if the video codec bit rate can impact user perception of asynchrony. The interaction between asynchrony and packet loss (audio or video) can lead to visual masking as well. Thus, we will give answer elements to this problematic that it is not yet studied in the literature.

We realized non-interactive subjective audiovisual tests with two objective of assessing audio, video and audiovisual qualities and defining asynchrony perception thresholds. Two separate tests have been conducted. One in laboratory, following the protocol of ITU-T P.911 [61], and the second one on a crowdsourcing platform. In fact, laboratory quality studies are time consuming and expensive, so researchers often run small studies with less coverage in terms of tested conditions. The crowdsourcing approach allows having a large and diverse panel of subjects in realistic user settings. Some researchers on QoE assessment developed specific crowdsourcing platforms and show the efficiency of the crowdsourcing method [167, 168, 169, 170].

In this chapter, we show the results of the two subjective tests conducted in order to better understand the influence of the time offset between the audio and the video media streams of videotelephony contents with the presence of other impairments. We also compare between the subjective perception of quality and asynchrony in laboratory and crowdsourcing contexts.

4.1 Test plan

Since most elements of the test plan are common between the two subjective studies, in this section only the differences will be highlighted. In our subjective tests we simulated asynchrony conditions with the presence of video IP packet loss, audio IP packet loss, video coding bitrate and video resolution. We simulated the video and

audio packet losses on the audiovisual sequences by applying a random loss pattern. The values of the packet loss percentages had been set empirically by experts who observed the results on the contents and selected those which represent an actual case of use. In Chapter 3 we showed that an audio IP packet loss value under 5% has no significance impact on subjective quality perception as a value over 5% is not realistic and the quality become very degraded. Thus, in our test plan we have only one condition for audio quality. We considered three video coding bit rate values in order to cover a wide range of visual quality (from good quality with 768 kbps to bad quality with 64 kbps). Three resolution conditions are studied to represent cases of use of a videoconferencing service on mobile, tablet and even PC.

The different values for all these variables are given in Table 4.1.

Audio/video delay (ms) (NOTE)	Video IP packet loss (%)	Audio IP packet loss (%)	Video coding bitrate (kbps)	Video resolution
-400	0	0	768	640 × 480 VGA
-300	1	5	384	320 × 240 QVGA
-250	2		64	1280 × 720 720p
0				
150				
300				
400				
NOTE negative values stand for when sound is delayed with respect to image, positive values stand for when image is delayed with respect to sound.				

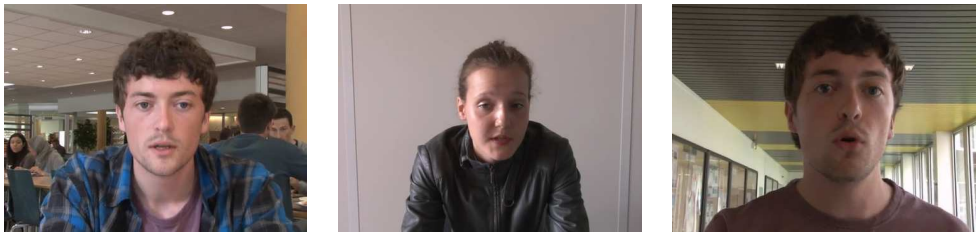
Table 4.1: Variables for the subjective test

Our reference condition consists in :

- no delay,
- no packet loss,
- bit rate at 768 kbps,
- resolution at 720p.

All seven conditions with different values of delay between audio and video have been repeated in presence of one single variation from the reference condition. There are in total seven possibilities (1% video IP packet loss, 2% video IP packet loss, 5% audio packet loss, 384 kbps, 64 kbps, VGA, QVGA). Together with the reference condition, this makes a total of $7 \times (7 + 1) = 56$ conditions.

For both tests, after each presentation the subjects were asked to evaluate the overall audiovisual quality (MOS_{AV}), the audio and video qualities (MOS_A^A and MOS_V^V) and the audio-video asynchrony (MOS_{synch})(See Appendix 9). To measure the perceived video, audio and audiovisual quality, we used an ACR five-level MOS scale and for the synchronization, we used a specific 5 point DMOS impairment scale as described in the recommendation P.911 [61].



Restaurant Scene 1 | Desk Scene 2 | Hall Scene 3

Figure 4.1: Screen shot of the used video contents

4.1.1 Tested contents

Three different video scenes have been selected, taken from our non-interactive subjective test that we described in Section 3.3. These contents are "Restaurant", "Desk" and "Hall" (see Figure 4.1). The duration of these scenes stand between 8 and 10 seconds. The selection of these three scenes has been driven by the need to have a wide representation of the bi-dimensional space composed of the dimensions “spatial information” and “temporal information” (see Figure 4.6) representing respectively the complexity and the amount of motion in the video part of the sequences, as defined in P.911.

Concerning the audio of these scenes, they represent also a good variety. The "Desk" scene is recorded in a quiet place, while the "Restaurant" scene includes some cafeteria noise and the "Hall" scene has a little reverberation. The audio signal has been coded with the same codec used in our previous tests which is AMR WB codec at 23.85 kbps.

In total, 56 conditions with 3 scenes bring to a total number of 168 sequences to view and assess.

4.1.2 Laboratory subjective test procedure

For the test in laboratory, 32 persons were involved (23 females and 9 males with ages from 16 to 55 years). To view and assess 168 sequences, the test for each participant took approximately two hours, divided into two one-hour sessions with a break. Before that, they had a training session on 5 sequences in order to become familiar with the test procedure and adjust the viewing distance and the sound volume.

The material to run the test was composed of:

- a PC screen (DELL 24”) where the video part of the sequences was displayed,
- a high quality headset (signature connected to an amplifier (STAX, SRM-006tII) in order to adjust the sound volume,
- a tablet to enter the answers to all four questions after each sequence.

The room where the test has been conducted is quiet and isolated from outside (acoustically treated); the light has been adjusted to 20 lux. All the materials were put on a table and the tester were sitting on a mobile chair allowing adjusting the viewing distance.

4.1.3 Crowdsourcing subjective test procedure

The concept of crowdsourcing, as a novel QoE assessment methodology, means to outsource subjective studies to a crowd in the Internet and calling on outsider testers to realize them. The crowdsourcing process consists in recruiting anonymous group of people to perform an audiovisual subjective quality test with their own devices and in their own environment. This unsupervised test context raises several difficulties and challenges [171]. In fact, crowdsourcing approach offers several facilities and advantages, allowing to benefit from a very large number of participants with a reduced cost compared to the standard laboratory test, as well as to face the problem of insufficiency of the data obtained by the classical methods.

In our context, as far as the crowdsourcing test is concerned, the test procedure has to be adjusted. In particular, due to the test duration constraints (less than 15 minutes), we chose to reduce by a factor of 6 the global corpus of 168 test sequences, so we decided that the number of scores that each participant visualized and assessed is only 28 sequences.

It is assumed in P.911 that the minimum number of participants to a subjective test is 15 in order to have good consistency and accuracy. Thus, the number of scores for all sequences when applying the crowdsourcing approach was equivalent or higher than this threshold.

A possibility could have been to design 6 separate tests comprising each of them 28 sequences, and to propose it to at least 15 testers. We decided not to do so, because there was a significant risk that the content of all individual tests could not be equivalent in terms of perceived quality, introducing thus a bias. Instead, it has been decided to have a fully randomized choice of sequences to be proposed to each tester.

However, this selection has a drawback: one is never sure if all sequences received enough scores. In other words, the minimum number of testers to get involved in a test in order to obtain at least a given number of scores for all sequences is unknown. In order to get some good idea, we ran 500 iterations of the selection of 28 elements in a whole set of 168 repeated 100, 120 or 150 times, and we looked at the number of these elements that were selected at least 15 times. The results are presented in Table 4.2.

This means that, if 150 persons are involved in the test, at least 163 sequences will be viewed and scores 15 times or more, and it is very likely that all of them

Number of testers	100	120	150
Minimum	109	143	163
Maximum	131	162	168

Table 4.2: Number of sequences with at least 15 scores

will. This is why our test has been designed for approximately this number of participants. At the end of the test, there was still the possibility to ask a few people to test a fixed set of sequences containing the ones with less than 15 scores. In this study all sequences have been viewed and scored at least 15 times as expected. For future studies on subjective quality test in crowdsourcing, we suggest to consider such a random approach. In our methodology we want to clarify that the number of testers per condition is considered as limited. We did not try to make a real crowdsourcing campaign but rather to use crowdsourcing tools to try to replace a formal subjective test.

The existing crowdsourcing frameworks for QoE assessment were studied and compared in [172]. In our case, the main criteria for our choice is that, since the audio part of the tested sequences is in French, native speakers for this language were required. Thus, we chose a crowdsourcing platform called "FouleFactory" [173] as it is the only platform that could allow this as far as we are aware. This platform responded well to our requirements, and in addition it assured a massive recruitment of testers (among their database of 50000 users) and it took less than 24 hours to get our 120 ratings. FouleFactory was in charge of rewarding the testers. Once recruited, testers were redirected to the URL of the test platform itself, on a server hosted by Orange.

A total of 146 persons took part in this test. 120 of them have been recruited by Foulefactory, the remaining were voluntary employees of Orange, not expert in quality assessment.

Once connected to the test platform, each tester had to respect the following protocol before starting the test:

- read a first page (see below Figure 4.2) giving information and simple recommendations concerning how to pass the test,
- visualize and evaluate five learning sequences (the same ones than for the P.911 test) in order to get more familiar with the impairments and with the scoring scales.

At the end of the test, the tester had to answer to a little questionnaire concerning some personal information useful for further statistics:

- What age group do you belong to?
    1. 18-30 years

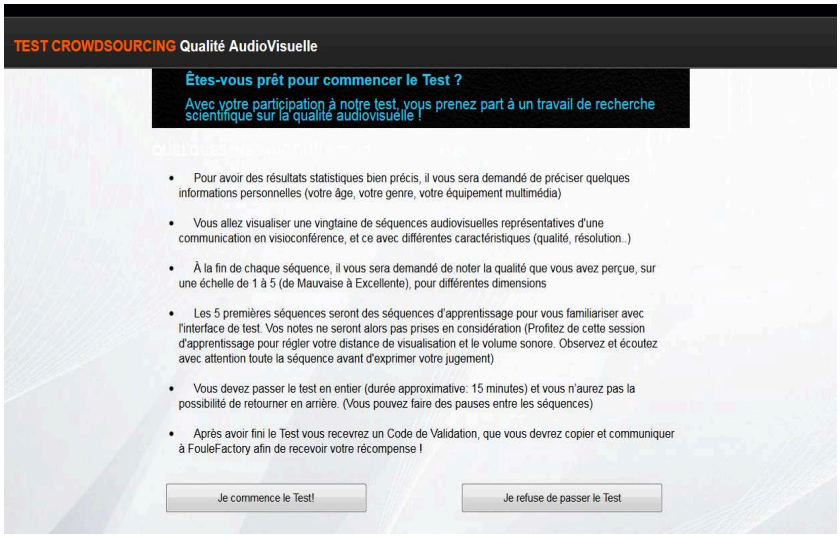


Figure 4.2: Recommendations before the test on the crowdsourcing platform (in French)labels of questions (in French)

2. 31-45 years
3. 46-60 years
4. more than 60 years

- Are you ... ?

1. Male
2. Female

- What type of audio equipment did you use for the test?

1. Headphone
2. Earphones
3. Loud speaker

Since the crowdsourcing environment is not controlled, researches proposed to add during and after the test, content questions and reliability checks in order to ensure the quality and the relevence of scores. These questions are considered to improve the reliability of the ratings by reinforcing the attention of workers and to be used to post-screening the possible unreliable workers. However, analysis results presented in [174] show that these consistency question are not efficient for post-screening and it is recommended to use the standard deviation as a criteria. Following these conclusions, in our test, no consistency question has been asked. However, before rewarding a tester (and taking his answers into account), the consistency of his answers was checked, by comparing them to the mean of all individual answers. Only the data from testers who gave at least for 5 conditions (out of 28) scores deviating by at least 1 point on the MOS scale from the mean value over all

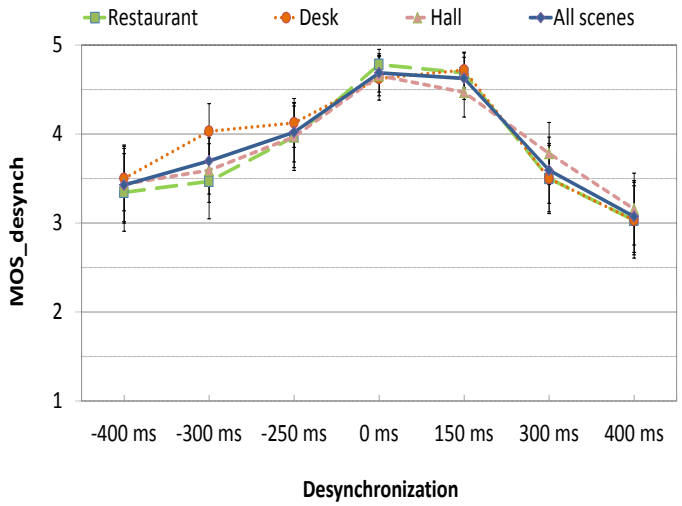


Figure 4.3: Perception of asynchrony in absence of other factors

NOTE 1: negative values stand for when sound is delayed with respect to image, positive values stand for when image is delayed with respect to sound; NOTE 2 : 1 = Very annoying ; 2 = Annoying ; 3 = Slightly annoying; 4 = Perceptible but not annoying; 5 = Imperceptible

testers for the four questions were discarded. This concerned 4 persons out of 120. A finer screening of scores has also been performed as usually done after formal subjective tests [156].

4.2 Subjective test analysis

The main reason to launch this series of tests was the need for a better knowledge of interaction between audio-video synchronization and other QoE factors. The answer to this question can mainly be found when analyzing the scores obtained by the "desynchronization" annoyance question. In this section, we examine this question for each factor. Before that, an examination of the answers to this question in the reference condition is necessary. They are illustrated on Figure 4.3.

Fig. 4.3 shows that the type of scene has no influence on the asynchrony perception in the reference condition, in this study. Furthermore, the acceptability thresholds are in line with previous knowledge [46, 166]:

- an image delayed by 150 ms is not perceived, but a sound delayed by 250 is perceived (but not annoying),
- the asynchrony is more perceptible for a given delay timing when audio leads video than when audio lags video.

In the following, we will see how far the introduction of video and audio quality factors can influence these observations.

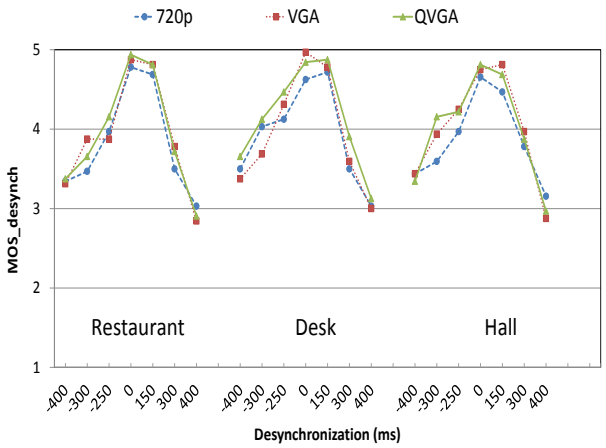


Figure 4.4: Influence of video resolution on the perception of asynchrony

4.2.1 Video resolution

The fact to have a smaller image is not a quality factor by itself, even if the results of the subjective tests show that the smallest images (with unchanged bitrate at 768 kbps) get better scores. This applies also for the question on asynchrony. On Figure 4.4, we can see indeed that this factor has small influence on the perception of asynchrony. This effect is identical for all levels of spatial complexity tested in this study..

4.2.2 Video coding bitrate

The results show that the perceived quality decreases with the video coding bitrate. This concerns not only pure video quality, but also global quality, as well as the perception of asynchrony. This decrease depends on the type of video content, and in particular its spatial complexity: “Desk“ (exhibiting the lowest spatial complexity) is the least concerned (a little bit more than “Restaurant“) while “Hall“ (exhibiting the highest spatial complexity) is the most impacted (see Figure Fig. 4.5). As far as the perception of asynchrony is concerned, it is influenced by video coding bit rate only at very low rates. This influence is especially visible for The “Hall“ scene, with scores dropping down close to 3 even without delay between audio and video (the global video quality is so bad that it is no longer possible to follow the movements of the lips), while for other scenes scores remain above 4 for low delays.

4.2.3 Video IP packet loss

Here again, the increase of the magnitude of this impairments results, without surprise, into a decrease in the asynchrony annoyance scales. This decrease depends on the scene, and again, “Hall“ scene, with the highest complexity, is the most impacted, as can be seen on Figure 4.6. An interesting thing to remark is that for all types of scene, with video delays of 300 ms or more, the perception of asynchrony

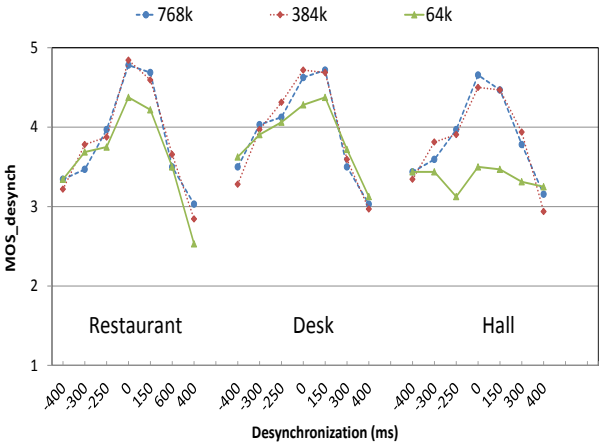


Figure 4.5: Influence of video bit rate on the perception of asynchrony

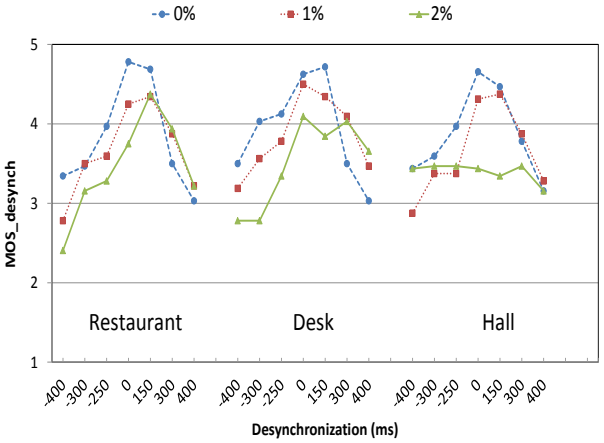


Figure 4.6: Influence of video IP packet loss on the perception of asynchrony

seems to decrease when packet loss is present (for "Restaurant" and "Desk") when it increases with audio delay.

4.2.4 Audio IP packet loss

From Figure 4.7) we notice that the scores for asynchrony perception reach lower values in the presence of audio packet loss. This decrease depends on the audio content of the tested scene, and in particular on the presence of noise in the background, which is the case for "Restaurant" (cafeteria noise). We see also the same trend as in the case of video packet loss when the video delay is high.

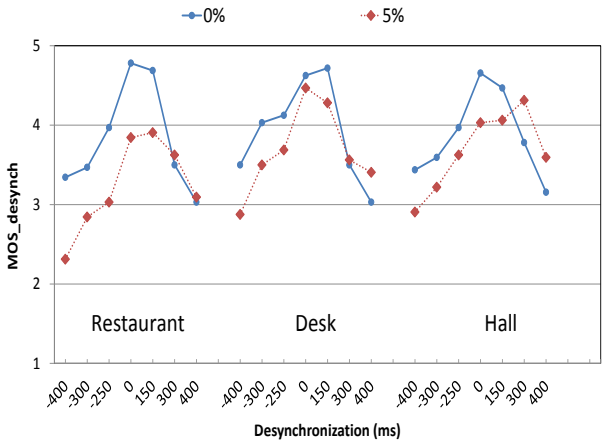


Figure 4.7: Influence of audio IP packet loss on the perception of asynchrony

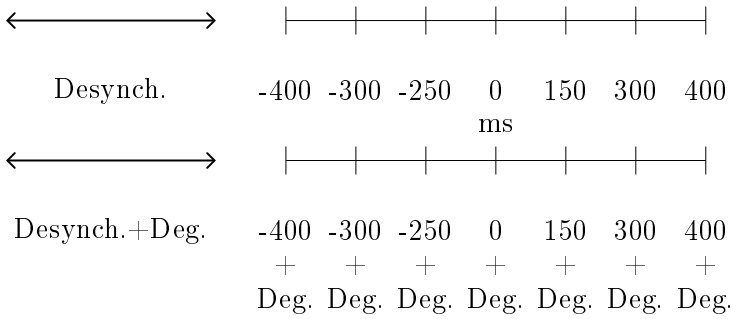


Figure 4.8: Blocks of test conditions. "Deg." is equivalent to: 1%VPL (Video Packet Loss),2%VPL, 5%APL (Audio Packet Loss), 384 kbps, 64 kbps, QVGA and VGA.

4.3 Comparison between results from laboratory and crowdsourcing tests

As seen in the section 4.1, both tests have been conducted with as much common characteristics as possible. The main differences are the number of scores per condition (32 in the P.911 test, between 15 and 25 for the crowdsourcing test), the number of sequences tested by each tester (all in the laboratory test, only 1/6th of them for the crowdsourcing test) and without the control of the testing conditions (lighting, viewing distance, screen, listening device) for the crowdsourcing test.

In Figures 4.9 to 4.13, one can see the mean scores for all 168 tested sequences, presented question by question for each content and for the mean over all the contents. The green curves show the scores for the laboratory test, the red ones for the crowdsourcing one. Conditions are divided into 8 blocks of 7 conditions. We present a zoom on the axis in Figure 4.8.

4.3.1 Global quality

The red and green curves in Figure 4.9 are rather close to each other. This corresponds to a very good level of correlation between both sets of data (around 92 %). However, it can be seen that video impairments (packet loss, lower bit rate) are scored more severely by testers following the crowdsourcing procedure (this is mostly visible for the "Restaurant" scene). This could be explained by the fact that the viewing distance (out of control) can be shorter for them than for formal tests in laboratory where a distance equivalent to 3 times the height of the screen has been applied by default.

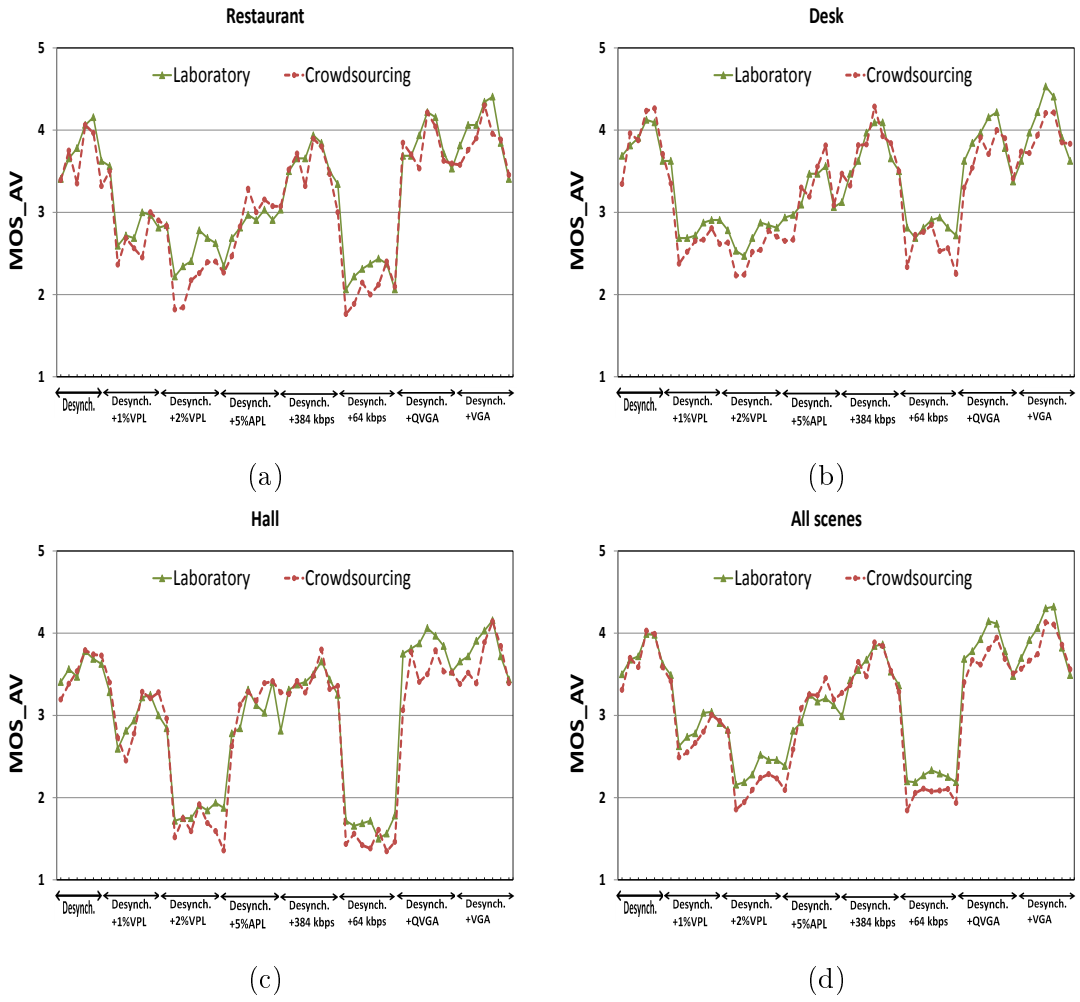


Figure 4.9: Comparison of mean scores of both tests for global quality

4.3.2 Audio quality

This is the question with the lowest level of correlation between both sets of data (around 60% only), as illustrated by Figure 4.10. In particular, laboratory results

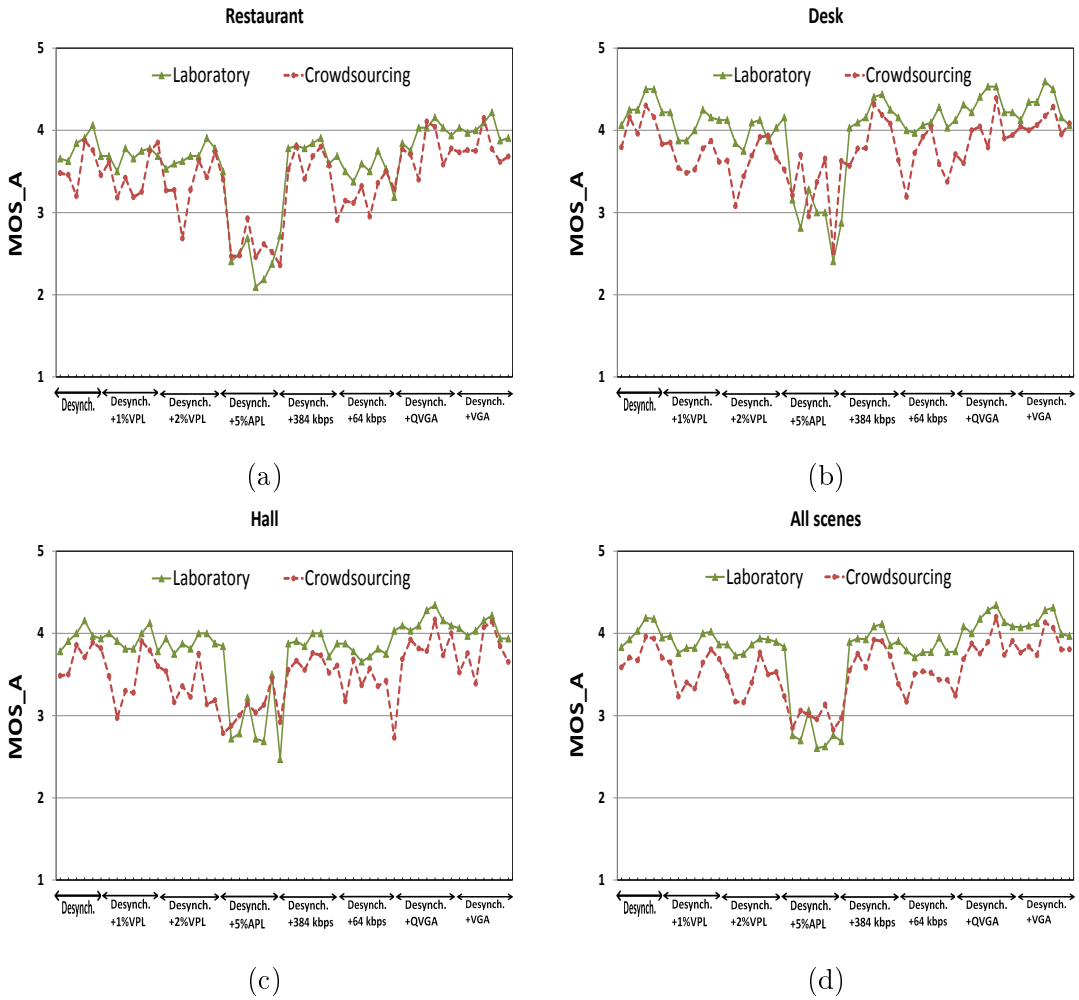


Figure 4.10: Comparison of mean scores of both tests for audio quality

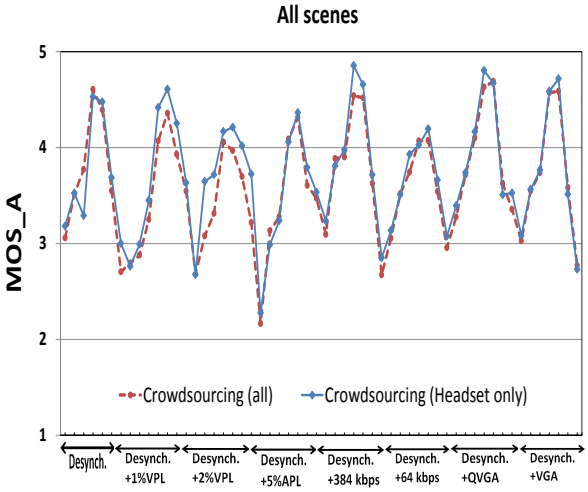


Figure 4.11: Comparison of mean scores of both tests for audio quality with a distinction based on listening device

show a very good discrimination between conditions with and without audio impairments, whereas this is not the case for the crowdsourcing test. This is particularly true with the "Hall" scene where the red curve does not exhibit lower scores for the conditions with audio packet loss compared to conditions with video packet loss or video coding with low bitrate.

The relatively low correlation resulting from this finding could be explained by the fact that the test environment was out of our control. Furthermore, most testers (86 out of 146) used loudspeakers. Figure 4.11 is similar to Figure 4.10, but the mean scores from the crowdsourcing testers using only headset have been added (blue curve). Unfortunately, the number of scores per condition becomes then too low (down to 5 for some sequences) to have fully relevant statistics, but a quick look at the relative positions of curves shows that the discrimination between conditions with and without audio impairments is enhanced with headset. This result is another good illustration of the difficulty to master audio listening conditions outside a laboratory environment.

4.3.3 Video quality

This is the question for which both sets of data are best correlated, up to 95% (see also Figure 4.12). Here, the replacement of a laboratory test by an approach based on crowdsourcing is obviously less problematic. This can be explained by the fact that difference in terms of media rendering is not as big as for audio between laboratory and home contexts for this study.

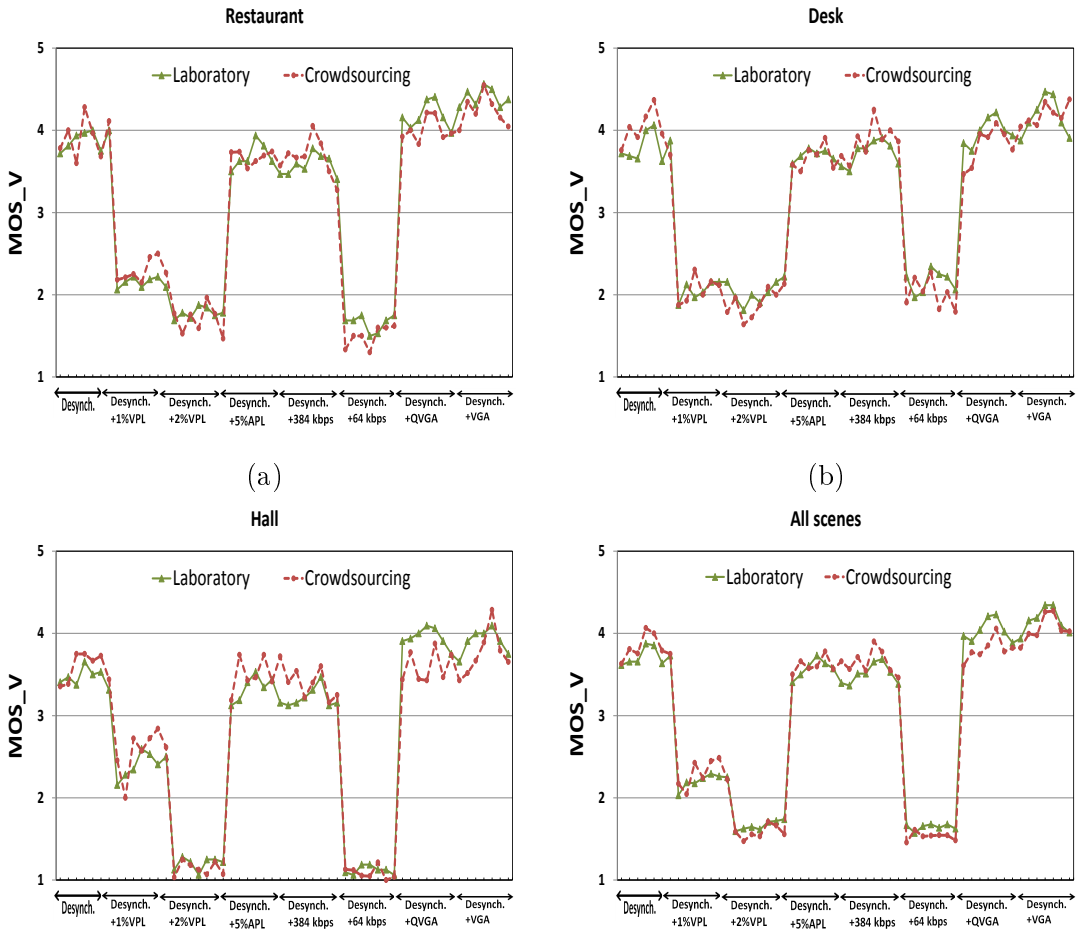


Figure 4.12: Comparison of mean scores of both tests for video quality

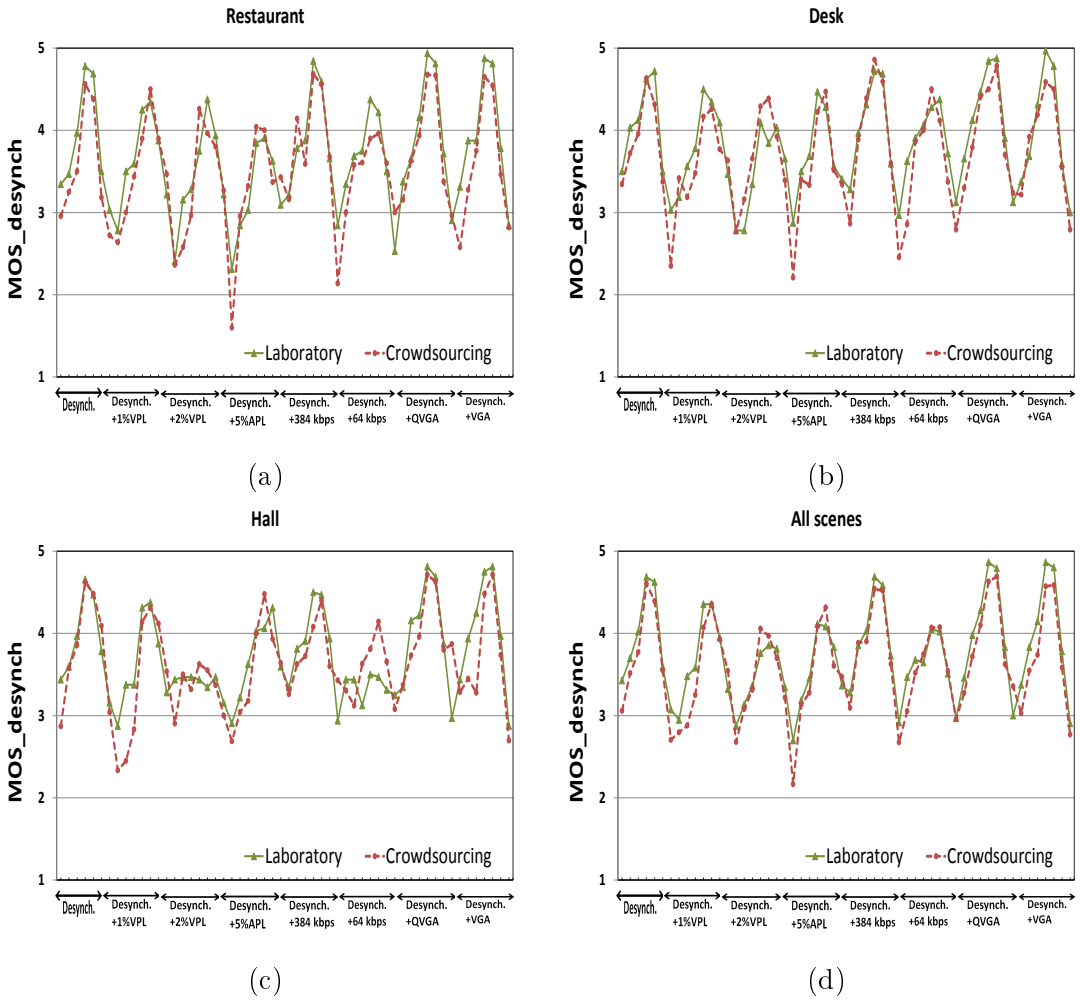


Figure 4.13: Comparison of mean scores of both tests for perception of asynchronism

4.3.4 Desynchronization perceptibility

For the desynchronization perceptibility, the correlation between both sets of data reaches 78%. As illustrated on Figure 4.13, generally, scores in the crowdsourcing context are more severe, in particular with the highest offsets between sound and image, but both methods are equivalent in terms of discrimination between good and bad conditions (i.e. the difference between scores with and without delay is large)

One can also see that both groups of testers judge the influence of video and audio quality factors on the perception of desynchronization more or less in the same way. The reason why the correlation is lower than for video and global quality questions is the same than for audio quality. If the sound is more difficult to listen to (as this is the case with the crowdsourcing approach), then sensitivity to an offset against image is certainly decreased.

An obvious factor affecting the delay between video and audio is the distance between the loudspeaker (or, more generally, the electro-acoustic transducer) and the tester. If reflections of the direct sound have a significant amplitude then these could provide misleading, or at least alternative, cues of synchronization. Headset provides a very good control in two ways: the delay introduced is very small and the ratio of direct to reverberant sound is very large. However, this is dependent on the goodness of fit. The fidelity of the loudspeaker, or the headset, also affects the character of the sound heard by the tester.

Some differences between scores with both test methods can however be observed in a few isolated cases. For instance, the participants in the laboratory test give lower score for all conditions with the “Hall“ scene and a 64 kbps video bitrate, whereas those following the crowdsourcing approach do not notice asynchrony problem with low delay.

As a global conclusion, we can say that the use of a crowdsourcing approach (with enough participants, allowing at least 15 scores per sequence under test) leads to results that are equivalent to those of laboratory P.911 tests when it comes to video quality and global audiovisual quality on various types of contents representative of a videotelephony conversation. Nevertheless, the results are less promising as far as perception of asynchronism and (mostly) audio quality is concerned, where the crowdsourcing approach yields underestimation of quality and lower discrimination between bad and good conditions. The difference between the media rendering hardware used in laboratory and at home, as well as the uncontrolled acoustic environment, is certainly the main explanation.

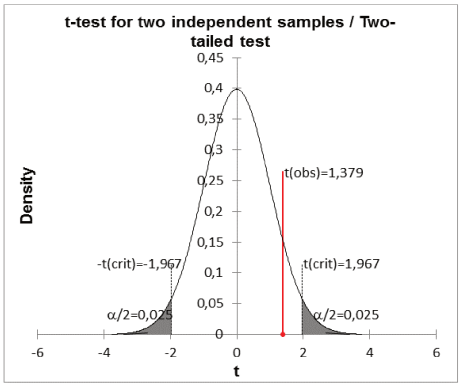


Figure 4.14: Distribution of T-values for global audiovisual quality

4.3.5 Statistical analysis of correlation

In order to compare the subjective scores given in the laboratory and crowdsourcing tests, we used the t-student parametric test. This test allows us to characterize if the difference between two samples are statistically significant or not.

We set the significant difference level to $\alpha = 5\%$ and the confidence interval to 95%. We consider the two hypothesis:

- H0: the difference between the averages is equal to 0
- Ha: the difference between the averages is different from 0

Global audiovisual quality
The statistical test calculates a p-value equal to 0.169 which is greater than the threshold significance level α. Thus, the null hypothesis H0 cannot be rejected. We confirm that there is not a significant difference between the audiovisual quality evaluation in a laboratory and in a crowdsourcing environment.

Audio quality
The found p-value is equal to 0.0001 which is lower than the threshold significance level. Therefore, the null hypothesis H0 must be rejected. Statistically, there is a significant difference between the audio quality evaluation in a laboratory and in a crowdsourcing environment.

Video quality
The obtained p-value is 0.881 which is greater than the threshold significance level. Thus, the null hypothesis H0 cannot be rejected. We confirm that there is not a significant difference between the video quality evaluation in a laboratory and in a crowdsourcing environment.

Synchronisation
The obtained p-value is 0.051 which is very close o the threshold significance level.

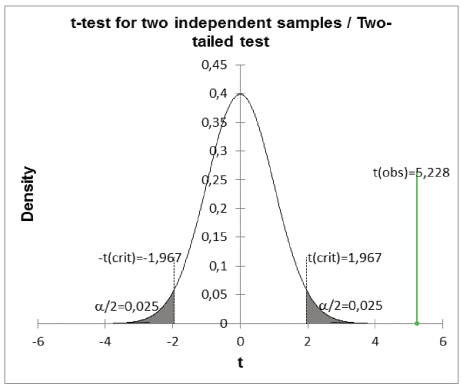


Figure 4.15: Distribution of T-values for audio quality

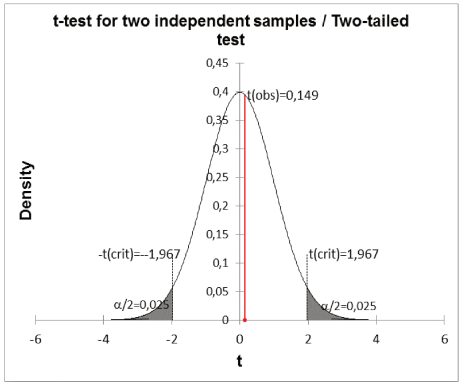


Figure 4.16: Distribution of T-values for video quality

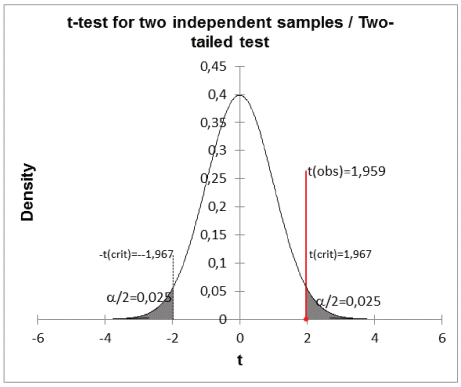


Figure 4.17: Distribution of T-values for synchronization perception

Thus, the null hypothesis H0 cannot be rejected. There is not a significant difference between the synchronization perception in a laboratory and in a crowdsourcing environment.

4.3.6 Outcomes

As a global outcome, we can say that the use of a crowdsourcing approach (with enough participants, allowing at least 15 scores per sequence under test) leads to results without significant difference compared with those of the laboratory test for the assessment of the video quality and global audiovisual quality on various types of contents representative of a videotelephony conversation. Nevertheless, the results are less promising as far as the desynchronization perceptibility and (mostly) the audio quality are concerned, where the crowdsourcing approach yields underestimation of the quality and lower discrimination between bad and good conditions. The difference between the material used in laboratory and at home, as well as the uncontrolled acoustic environment, are certainly the main reasons.

4.4 Conclusion and perspectives

The laboratory subjective test confirmed previous knowledge on the perception of desynchronization between sound and image in a conversational context. It brought however a few further interesting elements:

- spatial complexity of video contents and noisiness of audio contexts have a negative influence on the perception of desynchronization (the more complex or the more noisy, the worse in terms of perception),
- in presence of loss of audio of video information (resulting from IP packet loss), the desynchronization is less perceptible.

These elements of knowledge should be taken into account in the development

of all future subjective and objective QoE methods addressing the assessment of the perceived quality of audiovisual conversational services.

The comparison between the results of the laboratory test and those of the crowdsourcing test demonstrate that:

- If the test protocol used in the crowdsourcing test is faithful to the one in laboratory, it is possible to obtain a strong and statistically significant correlation to the scores given in the laboratory. This is true for the audiovisual, video and asynchrony perception questions. The only exception concerns the audio quality perception.
- The test design in the crowdsourcing context seems to have a minor influence on the reliability of subjective scores on these three scales. In particular, the restriction of each individual test to only 1/6th of the whole set of sequences is not an issue.
- In our context, the use of a consistency check based on content questions is not necessary. An a posteriori screening of scores is enough. This is also confirmed by outputs from similar studies conducted in audio-only contexts.
- The assessment of audio quality with a crowdsourcing approach seems more difficult in an audiovisual context than in a pure audio context. The uncontrolled environment in the test (used headset, loud speaker, volume adjustment, background noise ... etc.) has a greater impact. This implies also that the constraints in terms of listening conditions for future similar tests must be much stronger.

Chapter 5

Objective quality metrics evaluation

Contents

Introduction

Literature proposes numerous methods for objective evaluation of the audio, video and audiovisual qualities. The objective of this chapter is to evaluate the accuracy of the main existing approaches and metrics in predicting quality. The applications of objective quality evaluation are various. Post-processing, transmission, sensors or displays are elements that can be subject to specific quality criteria. Our main contribution is to investigate the performance of the objective models according to different impairments that can occur for instance in a video conference call.

5.1 Full reference video quality metrics

In this section, we evaluate the prediction accuracy of the full-reference video quality metrics defined in 2.4.2 on three different subjective databases: the Live Mobile video quality Database, the EPFL database and the videoconferencing database developed within the non-interactive test (3.3). In our work, we conduct an updated study of the existing set of full-reference metrics in the state of the art.

5.1.1 Performance evaluation and comparative study

We study the global full reference metrics in order to identify the representative metrics that have a good accuracy in predicting the subjective MOS score. As for all objective metrics, we evaluate the performance of the full reference metrics under study using three statistical indicators [175]:

1. Accuracy prediction: refers to the ability to predict the subjective quality ratings with low error. The Pearson Linear Correlation Coefficient (PLCC) was computed. For two datasets $X = \{x_1, x_2, ..., x_N\}$ and $Y = \{y_1, y_2, ..., y_N\}$ with $\overline{x}$ and $\overline{y}$ the means of the respective datasets, the PLCC is defined by:

$$PLCC = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2}\sqrt{\sum(y_i - \overline{y})^2}} \tag{5.1}$$

2. Monotonicity prediction: refers to the degree to which the relationship between the subjective quality ratings and the predicted measure can be described by a monotone function. The Spearman Rank Order Correlation Coefficient (SROCC) was used:

$$SROCC = \frac{\sum(X_i - X')(Y_i - Y')}{\sqrt{\sum(X_i - X')^2}\sqrt{\sum(Y_i - Y')^2}} \tag{5.2}$$

   with X_i and Y_i are the ranks of the ordered data series x_i and y_i respectively; X' and Y' denote the respective midranks.

3. Consistency prediction: measures the ratio of wrong predicted scores by the objective model to the total number of scores. The Root Mean Square Error (RMSE) was computed. For a dataset $\{x_1, x_2, ..., x_N\}$, with $\overline{x}$ is the mean value:

$$RMSE = \sqrt{\frac{1}{N}\sum(x_i - \overline{x})^2} \tag{5.3}$$

The PLCC and RMSE are computed after performing a non-linear mapping on the objective measures using the cubic polynomial mapping function recommended in [175]. This function is used in order to fit the objective model scores to the subjective scores.

The correlation values of the objective video quality metrics with the subjective scores are different. We cannot compare the performance of the metrics based only on the absolute difference between the correlations. It is necessary to investigate whether this difference in performance is statistically significant or not. For this purpose, we used a statistical Fisher test based on averaged quality scores as suggested in [175]. The F-test assumes that the scores are independent and have a Gaussian normal distribution. We used a Shapiro-Wilk normality test and we confirm that for all databases the data sets have the Gaussian distribution. We performed the F-test on the variance of the objective models at a 95% significance level.

The performance of all metrics in terms of the PLCC, SROCC and RMSE for the four video quality assessment databases are summarized on Table 5.1. The best performing metrics are highlighted in bold font for each test database and each criterion. It can be noted that the correlation scores on the EPFL database are significantly higher than the other databases and are in the range between 0.87% and 0.93%. This can be explained by the similarity of impairment types simulated in this database: only packet losses. It can be interpreted as an equivalent sensitivity of all the metrics to packet loss errors.

All three statistical measures (PLCC, SROCC and RMSE) show that generally three metrics, i.e. SSIMplus, ViS3 and VMAF outperform the other metrics. The common characteristic of these metrics is that they are video metrics ones that include the movement information in their quality assessment algorithms. On the other hand, classic image based metrics (PSNR, SSIM and MS-SSIM) are least correlated with the subjective video quality judgment.

By comparing the two VQM models (NTIA general and videoconferencing model) there is no significant difference between the correlation values for all databases except for Orange1 and Orange2 databases. For these sequences, VQM Videoconferencing model outperforms the NTIA General model. We can explain this result by the fact that the video contents of these databases are the closest to a videoconferencing context. Consequently, subjective scores are more influenced by this context. On the other hand, we note that both VQM General and VQM_V models have the less correlations on Orange2 database. This can be interpreted by the optimization of these models for video sequences encoded with H.263 and MPEG-4 [176], while Orange2 database contains H.256/HEVC encoded sequences.

The objective MOS prediction OPVQ model shows a good performance for EPFL and LIVE databases. Even though this model provides support for only a limited set of spatial resolutions (VGA, CIF and QCIF) and has been tested and validated for VGA resolution only, our correlation results prove that it could be applied on HD sequences. Furthermore, the coefficient parameters of the OPVQ model are trained on a data set containing quality impairments related to H. 264, H. 264 / SVC and MPEG - 4 coding, transmission errors, temporal dynamics (switches in video coding

bit rates during the sequence). We find all these degradations simulated in the Live Mobile database which explain the obtained correlation value 85%. However, the model has lower performances for the two Orange databases. We could explain that for two reasons: Orange databases contain 1) different degradation types (jitter, HEVC coding, frame rate changes) and 2) different contents from the training data set used to compute the mapping coefficients of the model.

The main strength of MOVIE algorithm is video quality estimation according to motion trajectories. The metric is accurate in detecting distortions that appear in regions containing movement. This explains the good MOVIE performance for EPFL and Orange 1 databases. In fact, it is known that unlike application distortions (coding, frame rate, resolution, etc.) independent from the content, transmission impairments (in particular the packet loss) infect objects on movement (which do not belong to the scene background).

A previous review [87] in 2011 showed that MOVIE had the best correlation with subjective opinions on LIVE video quality database, before the appearance of Vis3, SSIMplus and VMAF. The major drawback of MOVIE is its extremely high calculation complexity. MOVIE is the most complex metric in our experiment, which needs much more time than any other metric. This prevents its practical use in operational context.

Results shown in Table 5.1 reveal that Vis3 is competitive against the other metrics. The spatio-temporal dissimilarity estimation based on the video decomposition into spatio-temporal slices (STS) makes the algorithm less sensible to the temporal loss of alignment between the reference and the degraded sequences. In fact, due to the videoconferencing software and the recording process used to generate the Orange 1 database, we notice a slight misalignment in frames of the reference and those of the test videos. This difference impacts all the other objective metrics scores that are based on frame by frame comparison except ViS3 which is based on the Group Of Pictures (GOP) comparison. Thus, the most correlated metric for Orange 1 database (in terms of PLCC and SROCC) is ViS3. Furthermore, the performance comparison of ViS3 with the state-of-the-art video quality metrics in [97] reveals that for IP packet loss impairments, VQM General model and MOVIE outperform Vis3 for some databases. However, our correlation results on EPFL and Orange 1 databases prove that for videoconferencing contents ViS3 may be a good indicator for video quality in transmission error conditions too.

By comparing all the results we notice that generally, for all the databases and all degradation types, SSIMplus is one of the most competitive metrics. Despite the fact that our subjective test databases do not contain impairments in the range of device variability and viewing conditions, SSIMplus shows an accurate video quality prediction ability. In the results reported in Tab. 5.1, we precise that for the LIVE Mobile database we considered the SSIMplus metric values on all the sequences in-

cluding the frame freeze conditions. However, conditions with frozen frames have a large temporal misalignment between the reference and the degraded sequences which gives lower SSIMplus scores and thus decreases the correlation values. The SSIMplus software version that we used was not designed to handle freezing but there is a feature built in a commercial SSIMplus LIveMonitor software that automatically aligns frames up to 10 seconds difference.

Concerning the VMAF metric, it is highly correlated with the subjective results for all the databases except for the Orange1 database. We recall that the VMAF metric approach is based on a machine learning algorithm. Consequently its prediction accuracy largely depends on the characteristics of the training database: impairment types, codec configuration, resolution, frame rate, etc. Indeed, this model has been currently learned on sequences with only degradation caused by changes in resolution and different encoding bit rates. Thus, the poor correlation of VMAF for Orange1 database can be explained by the fact that only network impairments (packet loss and jitter) were simulated in this database. The EPFL database also contains only transmission errors but VMAF shows a good prediction accuracy (PLCC=91%, SROCC=92%, RMSE=0.55). In fact, IP network video packet loss depends highly on the used degradation simulator, the test bed and especially the video decoder and the jitter buffer. For the Orange1 database, some experts visualized the sequences and chose those with more perceived and annoying packet loss (degradation in regions of interest). Furthermore, a random model was used to simulate packet loss degradation for Orange1 database while the Gilbert-Elliot model was used for EPFL database. This difference between the models can explain the difference of the degradation perception.

Table 5.2 reports the statistical significance results of the F-test. Each entry in the table consists of 4 symbols corresponding to the databases "EPFL", "LIVE Mobile", "Orange1" and "Orange2". The symbol "+" indicates that the statistical performance of the VQA metric in the column is superior to that of the metric in the row. The symbol "-" means the opposite, while "0" indicates that the statistical performance of the metric in the row is equivalent to that of the metric in the column. Generally, statistical analysis shows that at a 95% confidence interval, all other metrics outperform PSNR and SSIM. It also proves that most consistent results with a high accuracy have been achieved by three metrics, i.e. ViS3, SSIMplus and VMAF.

	PSNR	SSIM	MS-SSIM	VQM-G	VQM-V	OPVQ	MOVIE	Vis3	SSIMplus	VMAF
EPFL database										
PLCC	0,88	0,89	0,89	0.90	0.89	0.91	0,87	**0,92**	**0,93**	**0,91**
SROCC	0,87	0,91	0,92	0.88	0.90	0.89	0,87	**0,90**	**0,92**	**0,92**
RMSE	0,68	0,66	0,65	0.61	0.65	0.60	0,71	**0,58**	**0,54**	**0,55**
Live Mobile database										
PLCC	0,71	0,65	0,65	0.83	0.82	**0.85**	0,71	0,84	**0.84**	**0,86**
SROCC	0,65	0,60	0,65	0.79	0.77	**0.82**	0,64	0,75	**0.76**	**0,77**
RMSE	0,62	0,66	0,66	0.50	0.52	**0.52**	0,61	0,52	**0.46**	**0,45**
Orange database 1										
PLCC	0,72	0,79	**0,81**	0.69	0.72	0.66	0,74	**0,85**	**0,79**	0,22
SROCC	0.68	0,71	**0,77**	0.72	0.74	0.67	0,72	**0,82**	**0,74**	0,23
RMSE	0.45	0.46	**0,46**	0.49	0.46	0.51	0,53	**0.42**	**0,48**	0,68
Orange database 2										
PLCC	0.48	0.52	0.48	0.55	0.58	0.57	0.73	**0.74**	**0.81**	**0.82**
SROCC	0.57	0.63	0.62	0.32	0.37	0.54	0.53	**0.91**	**0.75**	**0.76**
RMSE	0.61	0.60	0.61	0.53	0.51	0.61	0.54	**0.52**	**0.41**	**0.43**

Table 5.1: Statistical correlations of full reference metrics with the MOS scores

	PSNR	SSIM	MS-SSIM	VQM-G	VQM-V	OPVQ	MOVIE	ViS3	SSIMplus	VMAF
PSNR	0 0 0 0	0 0 0 0	0 0 + 0	0 + 0 +	0 + 0 +	0 + 0 +	0 0 0 +	0 + + +	0 + + +	+ + - +
SSIM	0 0 0 0	0 0 0 0	0 0 0 0	0 + 0 0	0 + 0 +	0 + 0 +	0 + 0 +	0 + + +	+ + 0 +	0 + - +
MS-SSIM	0 0 - 0	0 0 0 0	0 0 0 0	0 + - +	0 + - +	0 + - +	0 + 0 +	0 + + +	0 + 0 +	0 + - +
VQMG	0 - 0 -	0 - 0 0	0 - + -	0 0 0 0	0 0 0 0	0 0 0 0	0 - + +	0 0 + +	0 + + +	0 + - +
VQMV	0 - 0 -	0 - 0 -	0 - + -	0 0 0 0	0 0 0 0	0 0 - 0	0 - 0 +	0 0 + +	0 + + +	0 + - +
OPVQ	0 - 0 -	0 - 0 -	0 - + -	0 0 0 0	0 0 + 0	0 0 0 0	0 - + +	0 0 + +	0 0 + +	0 0 - +
MOVIE	0 0 0 -	0 - 0 -	0 - 0 -	0 + - -	0 + 0 -	0 + - -	0 0 0 0	+ + + +	+ + + +	+ + - +
ViS3	0 - - -	0 - - -	0 - - -	0 0 - -	0 0 - -	0 0 - -	- - - 0	0 0 0 0	0 0 - +	0 0 - +
SSIMplus	0 - - -	- - 0 -	0 - 0 -	0 0 - -	0 0 - -	0 0 - -	- - - -	0 0 + -	0 0 0 0	0 0 - 0
VMAF	- - + -	0 - + -	0 - + -	0 - + -	0 - + -	0 0 + -	- - + -	0 0 + -	0 0 + 0	0 0 0 0

Table 5.2: Statistical significance table based on residuals between model predictions and the MOS values for respectively the EPFL, LIVE Mobile, Orange1 and Orange2 databases. The symbol "+" indicates that the statistical performance of the VQA metric in the column is superior to the one in the row. The symbol "-" means the opposite, while "0" indicates that the statistical performance of the metrics in the row and in the column are equivalents.

5.1.2 Summary

In this section we have conducted an updated survey of the developed media-layer full reference objective video quality models. We carried out a performance comparison of ten different objective metrics in the context of video calling and videoconferencing. The comparison of metrics was performed based on their prediction accuracy, monotonicity and stability. In this study, we used two public video quality databases (EPFL and LIVE Mobile) and two databases created as part of our audiovisual videoconferencing subjective quality tests in Orange Labs. Experimental results show that metrics which include information about temporal video aspect in the quality estimation algorithm outperform other metrics. For the EPFL database which contains only the packet loss transmission errors, all the metrics are well correlated with the subjective video quality perception, with a little preference for OPVQ, Vis3, SSIMplus and VMAF. For the same degradation type with contents closer to those in the videoconferencing context, ViS3 statistically outperforms the other tested metrics.

In what concerns impairments caused by the H.264 and the HEVC coding bitrates VMAF and SSIMplus are the most competitive metrics. For a cross degradation types database, OPVQ, VMAF, ViS3 and SSIMplus have an equal statistical performance that exceed the other metrics. Thus, experimental results show that there is no universal metric which is best for all distortion types and contents. For evaluating the influence of codec type, coding bitrate and frame rate changes, OPVQ, ViS3, SSIMplus and VMAF can give out objective scores better correlated with the MOS. However, further studies are needed to optimize the OPVQ algorithm for the new generation of video codecs such as the HEVC. In the case of network transmission errors, we have a high probability to obtain a temporal misalignment between the reference and the degraded sequences. As a result, the scores of metrics based on frame by frame comparison are biased. In that case, we recommend the use of the ViS3 metric because its algorithm is based on computing quality on the GOP and the STS. VMAF is a promising model for video quality since it is constructed using the machine learning approach. Its performance can be enhanced by enriching the learning data set with large simple of impairment and contents types, and by training other better objective metrics such as the SSIMplus, ViS3...etc.

5.2 No reference video quality metrics

The majority of the state of the art studies about no reference metrics are limited to the common degradation types and are dedicated to a specific context (streaming, MPEG, HEVC coding, IP transmission, etc) [177, 178, 179, 180, 181]. The main focus of our work is automatic assessment of video quality in real time conversational services. In this context, it is necessary to detect a large set of distortion types. We consider no reference video metrics that have not been evaluated previously. These metrics are the key indicators of audiovisual quality developed by the De-

partment of Telecommunications in the AGH University of Science and Technology. This research work is a part of the MOAVI (Monitoring Of Audiovisual Quality by Key Indicators) project within the Video Quality Experts Group (VQEG)[2]. The proposed metrics estimate the presence of different video quality impairments such as Blockiness, Block loss, Blur, Noise, Flickering, etc. Our evaluation results aim to identify the conditions under which these simple NR metrics can be used effectively and in line with human perception in our use case for video-telephony.

In our study we consider also a global no reference video quality metric in order to compare its performance with the one of the single artifact based MOAVI metrics. We chose the completely blind Video Integrity Oracle VIIDEO metrics because its a video based metric (take into account the temporal aspect of the video) unlike other metrics that are image quality based.

5.2.1 Definition of MOAVI key indicators

By exploring end to end transmission of a video content in a multimedia conversation stream, the artefact Key Performance Indicators (KPI) can be grouped into four categories [182].

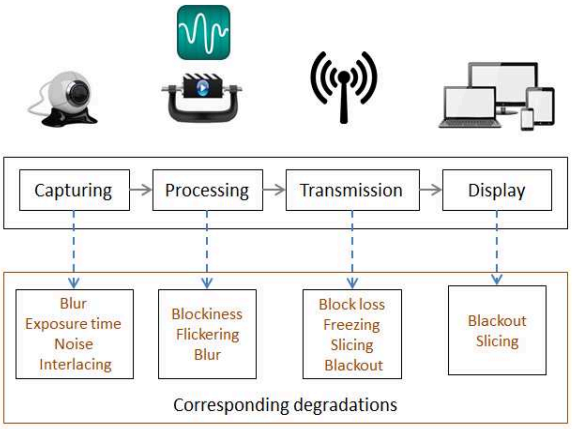


Figure 5.1: End-to-end transmission chain with the generated impairments

- Capturing : blur, exposure time, noise, interlacing.
- Processing: blockiness, flickering, blur.
- Transmission: blockloss, freezing, slicing, blackout.
- Display: blackout, slicing.

We selected a set of no reference metrics that we judge representative of the type of degradation that may infect a video conference or a video-telephony call.

Figure 5.2: Video indicators examples [2]

Therefore, in our study we put a focus on investigating the effect of the processing and transmission artifacts on the quality perceived by video services users.

Blockiness

It refers to the visibility of encoded blocks in the image. The implemented algorithm is descibed in [183]. Blockiness effect is detected by comparing separately the pixel luminance for intra and inter pairs of a single coding block. In order to consider the temporal aspect in a video and to be conform to a real use case application the metric is calculated on a time window (over all the video frames). The mean value for the window represents the blockiness level. The annoyance visibility threshold of Blockiness is equal to 0.9. Below this value, the artifact is more visible.

Blockloss

Block loss occurs when the packets containing the video stream are lost or damaged during transmission. This artifact is manifested by fixed color in regions of the image. This artifact is estimated by determining horizontal and vertical edges in every video frame. If these edges do not correspond to an object, the macro-block is classified as lost. The total number of lost events indicates the visibility of block loss artifact. The annoyance visibility threshold is equal to 5. Above this value lost locks are more visible.

Blur

This artifact is a deformation of the whole video frame, characterized by reduction of the sharpness in the contours and a loss of spatial details. The implemented algorithm is based on calculating the cosine of the angle between plane perpendiculars in adjacent pixels [183]. The annoyance visibility threshold is equal to 5. More the value of this metric is important, more the blur impairment is visible.

Flickering

It is a temporal artifact that appears mostly in the textured areas. It is illustrated by flicker of lines or blocks of frames, making the video unstable. Linked to block filtering in the decoder and in the encoder, the artifact is illustrated by strong difference in temporal contrast from one frame to another. The detection of this artifact is based on calculating the average absolute difference in pixel luminance for each 16×16 macro block [183]. Typical value for a sequence without distortion is equal to 0.125.

Freezing

Video freeze occurs when a picture is not updated. This distortion can be detected by checking for changes in the picture between consecutive decoded frames. A non zero value of this metric indicates the presence of frozen frames.

Slicing

Loss of some encoder video slices introduce high distortion to the video quality. Slicing artifact is manifested by destroyed video lines. Perception of this distortion is dependent on the encoder and decoder configuration. Slice prediction algorithms are implemented in decoder in order to reconstruct lost slices. Slicing upper value threshold for sequences without distortion is 0.

Spatial Activity

It describes the number of details on a video. A scene containing high frequencies corresponding to lot of details has a great spatial activity value. This metric is defined in the recommendation P.910 of the ITU [59] as the spatial information and is based on edge detection filtering. A sequence with normal SA has a value between 0 and 60. Above this threshold, a video is considered as spatially complex.

Temporal Activity

This metric indicates the amount of movements in a sequence. This metric is based on the motion difference feature which is the difference between the pixel values (of the luminance plane) at the same location but at successive times or frames. A sequence with normal TA has a value between 0 and 20. Greater the value of the metric, greater temporal activity is contained in the sequence.

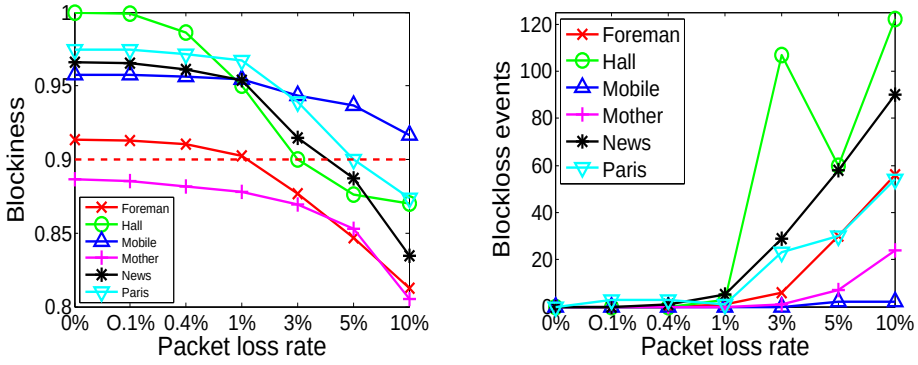
5.2.2 Performance evaluation and comparative study of MOAVI metrics

We evaluate the performance of the metrics under study using the same statistical indicators used in 5.1.1. As recommanded in [175] we applied a non-linear mapping before computing PLCC and RMSE coefficients. We used the cubic polynomial mapping function reported to perform well empirically.

In the web site of the metrics [2], a table with the annoyance visibility thresholds is set. Through our study, we find interesting to evaluate the accuracy of these thresholds and to investigate the variation of the metrics values according to the content and the degradation type.

	Blockiness	Blockloss	Blur	Freezing	Slicing	Flickering
PLCC	**0,73**	**0,57**	0,17	NA	0.15	0,34
SROCC	**0,75**	**0,63**	0,04	NA	0.15	0,07
RMSE	**0,91**	**1,10**	1,32	NA	1.32	1,36

Table 5.3: Statistical correlations of the non reference metrics with MOS scores of EPFL database



a b

Figure 5.3: Blockiness(a) and Blockloss events(b) variation on EPFL database

5.2.2.1 Evaluation on EPFL database

For the EPFL video quality database, only transmission errors are applied on source videos encoded in H.264/AVC. As shown in Table 5.3, the most correlated metrics with the MOS scores (and thus the most interesting in terms of diagnostics of perceived degradation due to IP impairments) are logically: Blockiness and Blockloss. This result is expected since these metrics correspond to the impairments caused by packet losses and bit errors.

For the Blockiness metric, a sequence without distortion has a value between 0.9 and 1.01. As represented in Fig. 5.3.a, from 3% of packet loss we notice the appearance of blockiness on the sequences "Foreman" and "Hall"; against from 5% for the sequences "News" and "Paris". However, for the "Mobile" sequence, even with 10% of packet losses the metric values are above the threshold of artifact detection. This can be explained by the fact that the "Mobile" sequence corresponds to the content whith the highest spatial and temporal activities (see Table 5.4). Therefore, consecutive frames on the video are different, which minimizes the detection of blocks. This result is in coherence with ones reported by P. Romaniak et al. in [183]. Based on the masking theory, they explained that high spatial and temporal activities are maskers to the blockiness artifact. On the other hand, "Mother" sequence has the lower SA and TA. Then, blockiness is visible even with 0% and 0.1% rates of packet losses.

	Foreman	**Hall**	**Mobile**	**Mother**	**News**	**Paris**
Spatial Activity	90.9	126.9	**206.5**	57.3	132.9	176.1
Temporal Activity	14.4	4.8	**22.6**	3.3	7.2	8.9

Table 5.4: Spatial and Temporal complexities of EPFL database

We can conclude that for transmission packet loss impairment, Blockiness metric is not independent and cannot be used in cross content assessment. It must be coupled with information on temporal and spatial activities. For the Blockloss metric, we compute for each content the total number of blockloss events. Actually, we scanned frame by frame the values of Blockloss metric, then we consider a Blockloss event when the value of the metric exceeds 5. The greater the total occurrence of events is, the greater the block loss impairment is visible and annoying. Results are shown in Fig. 5.3.b. For all the sequences, blockloss effect occur more from 3% of packet loss rate.

In order to identify for each condition the representative metric(s), we have applied a decision tree and regression algorithm on the objective and subjective scores with a significance level equal to 5%.

5.2.2.2 Evaluation on Live Mobile database

Table 5.5 shows that none of the selected no reference metrics is well correlated with the subjective MOS scores. Therefore, we cannot decide which are the more representative metrics for estimating the distortions in this database.

	Blockiness	Blockloss	Blur	Freezing	Slicing	Flickering
PLCC	0,29	0,17	0,16	0,22	0,15	0,32
SROCC	0,26	0,07	0,05	0,23	0,09	0,16
RMSE	0,84	0,86	0,86	0,85	0,86	0,83

Table 5.5: Statistical correlations of NR metrics with MOS scores of LIVE Mobile database

Since the Live Mobile video quality database contains sequences with different types of impairments and the metrics are distortion specific, we tried to evaluate the metrics by type of degradation. To do so, we divide the database according to the degradation types 3.5.1.

	Blockiness	Blockloss	Blur	Freezing	Slice	Flickering
Compression						
PLCC	**0,41**	0,24	**0,41**	NA	0,35	**0,60**
SROCC	**0,35**	0,18	**0.39**	NA	0,29	**0,59**
RMSE	**1,05**	1,11	**1,07**	NA	1,2	**0,91**
Frame freezes						
PLCC	0,25	**0,46**	0,26	**0,28**	0,23	0,21
SROCC	0,10	**0,36**	0,20	**0,20**	7,33E-04	0,07
RMSE	0,43	**0,39**	0,43	**0,43**	0,43	0,44
Rate adaptation						
PLCC	**0,35**	0,17	0,28	NA	0,27	**0,34**
SROCC	**0,44**	0,06	0,04	NA	0,07	**0,39**
RMSE	**0,61**	0,63	0,63	NA	0,63	**0,61**
Temporal dynamic						
PLCC	0,28	0,31	**0,36**	NA	0,23	0,23
SROCC	0,19	0,18	**0,20**	NA	0,18	0,19
RMSE	0,44	0,43	**0,42**	NA	0,44	0,44
Wireless channel packet loss						
PLCC	**0,49**	**0,60**	0,31	NA	**0,47**	0,43
SROCC	**0,47**	**0,57**	0,05	Na	**0,32**	0,31
RMSE	**0,97**	**0,89**	1,06	NA	**0,99**	0,97

Table 5.6: Correlation analysis for each condition of the LIVE database

	bf	hc	la	po	rb	sd	ss	tk
Spatial Activity	45,3	60,2	30,8	**90,9**	59,5	43,4	63,2	63,8
Temporal Activity	15,7	15,6	13,9	**22,**4	21,4	13,9	19,9	16,5

Table 5.7: Spatial and Temporal complexities of LIVE Mobile database

5.2.2.3 Compression

By analyzing the results on Table 5.6 we notice that the metrics that may be able to detect the compression artifacts are: Flickering, Blur and Blockiness. We find the same artifacts classified in the processing level (see Subsection 5.2.1). Previews study [183] shows that flickering is the most annoying temporal impairment due to inter-frames coding and in particular for H.264/AVC encoded sequences. We confirm this observation with the percentage of correlation equal to 60% between the flickering metric and the subjective scores.

We investigate the evolution of the most correlated metrics' values in order to evaluate the effectiveness of their annoyance visibility threshold values. For the Flickering and Blockiness metrics, all the results values are above the detection threshold which explains that these artifacts are visible in all the sequences. We confirm this results after visualizing the sequences.

Concerning the Blur metric, we obtained the lowest values (inferior to 5: the limit value for sequences without distortion) for the content "po". However, the sequence "la" had the greater values of Blur. In order to explain these results, we must consider the properties of video content. Thus, we must take into account the spatial and temporal complexities of the source sequences (see Table 5.7).

Greater spatial and temporal complexities lead to the non-detection of blur distortion. After applying a decision tree algorithm, we can conclude that for compression conditions Blockiness, Blur and Flickering metrics are to be considered.

5.2.2.4 Frame freezes

In the case of freezed frames, the subjects view a fixed image during few seconds. As a result, we find the best correlation with the metrics freezing and blockloss. Comparing to the other conditions we notice that the freezing metric results are non-zero only for this condition. As a sequence, we can confirm that this metric is able to detect the presence of image freeze. We consider the Freezing metric as the representative indicator in condition of frame freeze.

Moreover, since frame freeze is a transmission impairment, it is associated with the generation of blockloss in some frames. After analyzing the results of blockloss event indicator, we found that the values are non null only for the contents "la", "bf" and "sd" which correspond to the lower temporal ans spatial activities. This result is not sufficient to consider blockloss metric for freezing condition.

5.2.2.5 Rate adaptation

An interesting observation from the results is that single and abrupt switch from rate Rx to rate Ry and then switch back to Ry (where $Rx < Ry$) causes blockiness artifacts in the video. This impairment is visible in all the sequences.

5.2.2.6 Temporal Dynamics

Unlike the conditions of rate adaptation, here we have a multiple switches between bit rates: change from Rx to Ry with passing by an intermediate rate Rz. We note that the Blur metric is the more correlated with subjective quality perception. We can conclude that multiple rate switches cause the generation of blur in video sequences. Blur degradation is more visible on "la" content and less visible on "po" sequence.

5.2.2.7 Wireless channel packet loss

Simulating packet losses in the wireless channel is the source of several artifacts as shown in the table 5.6. We find the same metrics that we retain in the case of the EPFL database, as we have the same impairment types. Blockiness and Flickering are the more visible artifacts detected in the condition of wireless packet loss.

5.2.2.8 Evaluation on Orange videoconferencing dataset

Slicing, Flickering and Blockiness are the distortion specific metrics that characterize the IP packet loss impairments simulated on Orange video conference database. By examining the results of the metric Blockiness we observe that the values are quasi constant and in the range of sequences without distortion. Concerning the values of blockloss events, they are always equal to 0. After observing the sequences we notice the presence of artifacts associated to slicing distortion more than Blockiness distortion. Thus we represent in Fig. 5.4 the results of the metric slicing. The value of slicing metric increases with the rate of packet loss. As it is represented, the sequences "Park", "Hall" and "Poster" have the greater values for the metric. These scenes are highly temporal complex and this explains why slicing distortion is more visible.

	Blockiness	**Blockloss**	**Blur**	**Freezing**	**Slice**	**Flickering**
PLCC	**0,26**	0,25	0,20	NA	**0,42**	0,08
SROCC	**0,28**	0,26	0,14	NA	**0,39**	0,10
RMSE	**0,70**	0,70	0,71	NA	**0,66**	0,73

Table 5.8: Correlations of non reference metrics with MOS scores of Orange database

5.2.3 Completely Blind Video Integrity Oracle VIIDEO metric

VIIDEO [184] is a completely blind video quality metric which does not require the presence of the reference video or human judgments for training. The metric does not model any distortion specific information, but only models the statistical 'naturalness' (or lack thereof) of the video. The algorithm is based on the inter

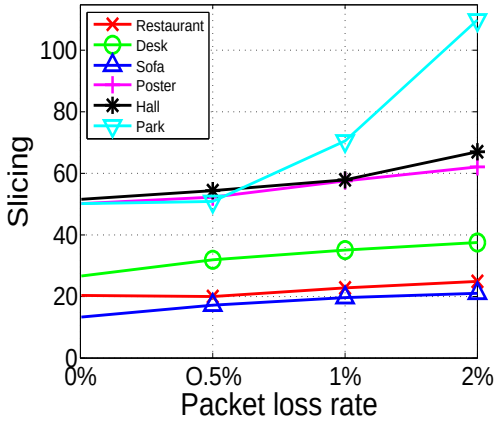


Figure 5.4: Slicing metric Orange database

sub band correlations to quantify the degree of distortion present in the video and hence to predict human judgments of video quality. Furthermore, the time complexity of every step in the video intrinsic integrity and distortion evaluation algorithm are analyzed. VIIDEO metric assumes that for a video of good quality, its local statistics of frame differences processed by local mean removal and divisive contrast normalization should follow a generalized Gaussian distribution.

We calculate the correlation between the VIIDEO quality values and the subjective MOS scores of the databases EPFL, LIVA and our subjective Orange databases. The results presented in Table 5.9 show that the VIIDEO metric outperforms the MOAVI single artifact based metrics. This can be explained by the fact that VIIDEO is a global quality estimation methods and it is more correlated with the MOS score which is also global subjective perception of the quality.

Database	PLCC	SROCC	RMSE
EPFL	0.8740	0.8434	0.7005
LIVE	0.6847	0.7180	0.6717
Orange	0.6725	0.6109	0.6688

Table 5.9: Correlations of VIIDEO non reference metric with MOS scores

5.2.4 Summary

In this section, we presented a performance evaluation study of six video quality assessment metrics developed by MOAVI VQEG project. The study involved three test databases with large sample of impairment types. We find that the metrics may be representative indicators of video quality. For each condition (encoding,

packet loss, signal attenuation, etc) we identified the representative metrics that we recommend to take into account (see Table 5.10). According to the obtained results, it can be seen that for transmission impairments, distortions perceived by end user can be manifested by block loss events, slicing or freezing. In what concerns impairments related to encoding, they are essentially blur, blockiness and flickering. These metrics and thresholds constitute a part of the tool box to diagnose video quality in communication services.

Distortion type	**Representative metrics**	**Threshold**
H.264 encoding	Blockiness Blur Flickering	0.9 5 0.125
Packet loss	Blockloss events slicing	19 68
Frame freeze	freezing Temporal Activity	0 20
Rate adaptation	Blockiness flickering	0.9 0.125
Temporal dynamics	Blur Spatial activity	5 60

Table 5.10: Summary of representative metrics for each condition

5.3 Audio quality metrics

In this section for the evaluation of objective audio quality models we will consider the POLQA model in SWB mode. The reasons why we study this model are:

1. it is representative of the first objective models able to characterize the perceived defects in the super wide band telephony communication context,
2. its code is accessible to us,
3. it is widely used in state of the art.

Other models exits, like PESQ [185] and the E-model [126], but none of them can be applied on SWB signals. This explains why we restrict our study on POLQA.We are investigating the predictive accuracy of this model in the case of a conversation audio recordings.

The database we considered here is the one of our non-interactive subjective test composed of 6 source sequences. Since the POLQA model is a full-reference model it was not possible to apply it to our interactive subjective test records since we do

Sequence	Restaurant	Desk	Sofa	Poster	Hall	Park
Pearson correlation	0.962	0.986	0.940	0.816	0.979	0.710
Spearman correlation	0.923	0.954	0.912	0.820	0.948	0.796
RMSE	0.15	0.13	0.11	0.16	0.18	0.25

Table 5.11: POLQA correlation with subjective scores

not have the corresponding reference signal.

We recall the audio degradation conditions applied in our database: 2%, 5% and 20% packet loss and 30ms of jitter. POLQA in its SWB mode provides an overall quality score ranging from 1 to 4.75. The correlation results of POLQA scores with subjective scores are shown in Table 5.11.

These results confirm the relevance of the POLQA model for assessing audio quality. Unfortunately, this tool is applicable only in a Full reference context. For an application in a SWB No-Reference context, there is currently no tool, but it is expected that soon ITU-T will standardize such a model (current work ongoing under the so-called P.SPELQ study item at ITU-T Q.9/12), with expected performance equivalent to POLQA [186].

5.4 Global audiovisual quality model: ITU-T G.1070 standard

In this section, we study the prediction accuracy and the relevance of the ITU-T Recommendation G.1070 "Opinion model for video-telephony applications" (2012) model [67] (including the recent proposed updates not yet included in the standard), initially meant for planning purposes only.

5.4.1 Performance study

As we showed in 2.4.1 most of the research studies for evaluating and enhancing the G.1070 model are only related to the video quality module. The global audiovisual quality estimated by the model including audio quality has not been investigated yet. An essential factor influencing the audiovisual quality of video phony applications is the synchronization between the audio and the video streams.

On another hand, the speech quality estimation of the G.1070 model is based on the ITU-T Recommendation G.107.1, known as the E-Model. Some studies within the SG12 of the ITU contribute to the development of the E-model. They show that the E-model is validated and largely accepted although there are some aspects under study such as delay, echo, additivity of equipment degradation factors, etc. Indeed, the subjects did not rate transmission delays as low as the E-model predicts. The present E-Model supports Wide Band audio signal and not yet Super

Wide Band signal. Furthermore, it is a planning model and it is not proved that it can be applicable for quality measure.

In the followings, we evaluate the G.1070 model based on our interactive and non-interactive subjective test results. The audio, video and audiovisual quality models are evaluated on our subjective tests databases using three performance metrics: the Pearson Linear Correlation Coefficient (PLCC), the Spearman Rank Order Correlation Coefficient (SROCC) and the Root Mean Square Error (RMSE) 5.1.1. The results are summarized in Table 5.12.

		PLCC	SROCC	RMSE
Non-interactive subjective test data base	Audiovisual quality	0.47	0.49	0.63
	Video quality	0.93	0.93	0.57
	Audio quality	0.12	0.43	1.42
Interactive subjective test data base	Audiovisual quality	0.36	0.51	0.76
	Video quality	0.85	0.73	0.41
	Audio quality	0.42	0.58	1.28

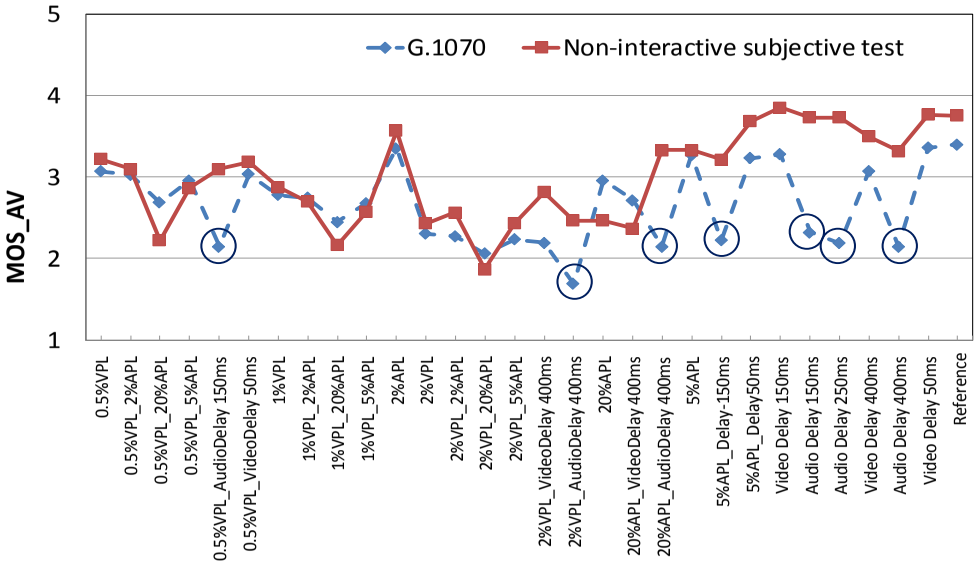
Table 5.12: G.1070 model correlation with subjective results

As being observed in Table 5.12 and Fig. 5.5, the audiovisual and audio modules have much lower performances compared to the video module. This result can be explained by the fact that the audio and audiovisual modules take as input parameters the speech and the video delays, whereas the video module does not. From Figure 5.5 we notice that all the conditions where the error between the G.1070 output and the subjective score is important, are the conditions with a speech delay (points circled). Thus, we can point out that G.1070 model underestimates the audio and audiovisual quality in cases of audio delay and it considers that this impairment deteriorates the quality with a greater extent than that perceived by subjects. If we ignore the audio delay conditions and we calculate the correlation between the model metric and the subjective scores we find the results presented in table 5.13.

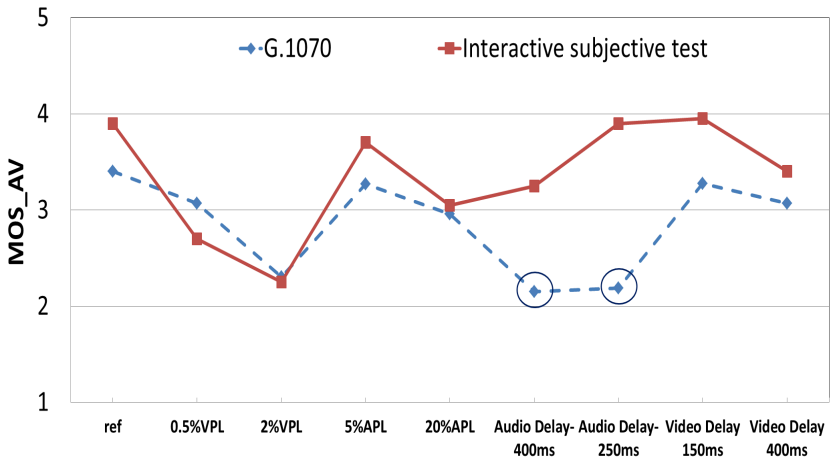
		PLCC	SROCC	RMSE
Non-interactive subjective test data base	Audiovisual quality	0.85	0.49	0.63
	Video quality	0.91	0.93	0.57
	Audio quality	0.98	0.91	1.42

Table 5.13: Correlation between G.1070 model results and subjective scores without audio delay conditions

Comparing with the correlation results in table 5.12, it is clear that this model provides a good estimation of subjective quality dealing with packet loss and video delay. For the non-interactive and the interactive subjective databases, we have the



(a)



(b)

Figure 5.5: Subjective results vs. G.1070 quality estimation in non-interactive (a) and interactive (b) contexts in conditions of Video Packet Loss (VPL), Audio Packet Loss (APL) and auido/video delay.

same results. This can indicate that this type of test scenario does not have an effect on the quality estimation process.

5.4.2 Proposal to enhance G.1070 model

In the previous subsection, we showed that the performance of ITU-T Recommendation G.1070 in terms of predicting the audio, video and audiovisual perceived qualities of video telephony communications was depending on whether audio or video was advanced in time compared to the other medium. When audio is delayed, the performance of the audio and audiovisual quality prediction (expressed in terms of ability to predict subjective test results) drops dramatically.

In the following, we will try to find explanations, and we will propose some possible corrections of the code of G.1070 in order to overcome this issue. It has to be mentioned however that the lack of data (only two conditions with audio delayed compared to video in an interactive context) cannot lead to firm conclusions, further data must be gathered and analyzed.

In the algorithms of G.1070's opinion model, the audio delay is taken into account at three locations in the calculations.

1. In the audio module, the computation of speech transmission rating Q is composed of one part without impact of delay ($Ie{-}eff$) and another one supposed to address the impact of talker echo ($Idte$). The formulate corresponding to this latter are modeling the annoyance due to talker echo, as per ITU-T G.131, based on two parameters: the echo loudness and the echo delay, making the assumption that the echo delay is equal to twice the one-way transmission delay.

2. In the audiovisual module, the global quality estimation MMq is a combination of two factors :

   (a) MMSV represents audio-visual quality and is itself a combination of video quality Vq and audio quality Sq. Sq is a translation of Q from the transmission rating scale to the MOS scale, thus it takes into account the talker echo factor $Idte$.

   (b) MMT is for the global impact of delay. It takes into account the absolute delay of the global stream (AD) plus the asynchronism between audio and video (MS).

The cause of the bad prediction can be found in either of these three sections of the G.1070 model's algorithm. We will see in the following how this can be checked for each of the potential causes, and what are the results once a modified algorithm is applied on the data used in 5.4.1. In order to test the correlation between the subjective MOS scores and the model results, we used three statistical

indicators: the Pearson Linear Correlation Coefficient (PLCC), the Spearman Rank Order Correlation Coefficient (SROCC) and the Root Mean Square Error (RMSE).

In the audio module

	with $Idte$			without $Idte$		
	PLCC	SROCC	RMSE	PLCC	SROCC	RMSE
Audio quality	0.12	0.43	1.42	0.92	0.87	0.78
Audiovisual global quality	0.47	0.49	0.63	0.84	0.89	0.34

Table 5.14: Compared performances of prediction by G.1070 of audiovisual quality scores with and without $Idte$ in the audio module for non-interactive subjective test

	with $Idte$			without $Idte$		
	PLCC	SROCC	RMSE	PLCC	SROCC	RMSE
Audio quality	0.42	0.58	1.23	0.74	0.09	0.99
Audiovisual global quality	0.36	0.51	0.76	0.87	0.89	0.42

Table 5.15: Compared performances of prediction by G.1070 of audiovisual quality scores with and without $Idte$ in the audio module for interactive subjective test

	with $Idte$			with Idd		
	PLCC	SROCC	RMSE	PLCC	SROCC	RMSE
Audio quality	0.12	0.43	1.42	0.77	0.82	0.78
Audiovisual global quality	0.47	0.49	0.63	0.81	0.85	0.37

Table 5.16: Compared performances of prediction by G.1070 of audiovisual quality scores with $Idte$ and Idd in the audio module for non-interactive subjective test

	with $Idte$			with Idd		
	PLCC	SROCC	RMSE	PLCC	SROCC	RMSE
Audio quality	0.42	0.58	1.23	0.81	0.52	0.89
Audiovisual global quality	0.36	0.51	0.76	0.86	0.88	0.44

Table 5.17: Compared performances of prediction by G.1070 of audiovisual quality scores with $Idte$ and Idd in the audio module for interactive subjective test

Depending on the echo level, the greater the audio delay is, the bigger $Idte$ gets and the smaller Q is. The simplest to check the relation between the delay and Q is well taken into account by the algorithm is to remove the computation of $Idte$ from the model. Thus, $Q = 93 - Ie - eff$. By doing so, there is no longer possibility to take audio delay into account in the computation of audio quality, so the expected result would be to have no really better prediction of audio quality in conditions where there is important delay.

However, the observed results are rather different.

- For the non-interactive context (see Table 5.14), audio and audiovisual global quality predictions are enhanced (more in terms of correlation than of mean error).
- For the interactive context (see Table 5.15), the trend is similar, with the notable exception of SROCC for the audio quality metric. This can be explained by the very small number of considered points, here a small change in score rank ordering can have a big impact on monotony measurement.

All this tends to prove that, during our test, even for interactive tests with high levels of asynchronism, subjects did not consider delay as a major matter of concern compared to other degradations (in our case: IP packet loss). A similar study on another database (with higher interactivity) seems necessary.

Thus, we can observe that the simple suppression of a factor is not satisfying. The effect of delay has to be taken into account somehow inside the audio module of G.1070, even for an application in contexts where this factor seems to play a minor role.

Since *Idte* is not giving full satisfaction, another solution has to be found. We did not investigate so far in our research, but we simply took a look at the source of the audio part of G.1070: the E-model of ITU-T Recommendation G.107. There, one can find a specific factor for pure delay, not present in G.1070. This factor is called *Idd*. By replacing the computation of *Idte* by the one of *Idd*, one can expect much more accurate results for the audio quality. This is proven at least on our data bases. As far as audiovisual quality is concerned, the improvement is also obvious as can be seen in Tables 5.16 and 5.17. Here again, the only observed exception concerns the SROCC for the audio quality question.

In the MMsv part of the audiovisual module

The formula between the audio, video and audiovisual quality combines them globally, without distinction between quality dimensions like delay. Therefore, we felt undesirable to modify it unless absolutely necessary. Since we found another way to enhance significantly the performance of the model, such a modification has not been undertaken.

In the MMt part of the audiovisual module

As seen in the section above on audio quality estimation, there are two potential ways to take pure audio delay into account in G.1070: in the audio module with the *Idd* factor, or in the audiovisual module with the computation of MS. We wondered whether both could be used together or if they could introduce some redundancy.

	with *Idd* only			with MS only			with both		
	PLCC	SROCC	RMSE	PLCC	SROCC	RMSE	PLCC	SROCC	RMSE
Audiovisual global quality	0.84	0.87	0.33	0.84	0.89	0.34	0.81	0.85	0.37

Table 5.18: Compared performances of prediction by G.1070 of audiovisual quality scores with *Idd* and MS (and both) in the audiovisual module for non-interactive subjective test

	with *Idd* only			with MS only			with both		
	PLCC	SROCC	RMSE	PLCC	SROCC	RMSE	PLCC	SROCC	RMSE
Audiovisual global quality	0.89	0.91	0.39	0.88	0.89	0.42	0.86	0.88	0.44

Table 5.19: Compared performances of prediction by G.1070 of audiovisual quality scores with *Idd* and MS (and both) in the audiovisual module for interactive subjective test

The results show that this is the case, both correlation factors and mean error are getting a little bit worse in case of joint use (see Tables 5.18 and 5.19). They prove also that there is no obvious best solution to take pure delay into account between *Idd* (with maybe a small advantage for the latter) and MS.

Discussion

The use of the talker echo factor *Idte* in G.1070 is clearly the major source for bad predictions when it comes to conditions with high audio delays (even if this needs to be studied further on a database where delay is much more felt as an issue by testers, and on a larger set of conditions). We recommend modifying G.1070 in order to remove this factor, or at least to recommend clearly a null default value for this factor (or a TELR default value above 100 dB) for videotelephony applications, where headphones are of wide use.

However, this removal has to be compensated by another factor to take into account the delay in the audio module of G.1070. For this purpose, we recommend to simply adopt the *Idd* factor from G.107. Nevertheless, by introducing this new factor, one generates redundancy with the MS factor used in the integration module of G.1070. Since the use of either MS or *Idd* seems to reach very similar results and performance, we recommend to get rid of this MS factor in cases where the audio delay is superior to video delay.

We discussed this issue on the basis of the contribution proposed by Orange at the joint session of Qs 7 and 13. It was raised that it is preferable to not introduce directly on the audio quality module a factor taking into account audio delay because otherwise it must do the same for the video module. The delay impact of quality

must be taken into account only in the multimedia quality integration function. Thus, MS must not be modified and the idea of introducing *Idd* from G.107 in the code of G.1070 must be abandoned.

On another hand, the case of audio quality assessment in the absence of echo is not well covered by the present model, and this is due to the fact that we can not put the *Idte* value to 0 because of a too weak default value of the attenuation of echo (65 db). It will therefore be allowed, in the particular case where we have the certainty of no echo to set *Idte* to 0.

5.4.3 Evaluation of the G.1070 extension

As presented in Chapter 2, Huawei Technologies Co. Ltd. proposed an extension of the G.1070 model to take into consideration the H.264 codec in its High Profile (HP) and Baseline Profile (BP) with different parameters. We are interested in evaluating this proposition, in order to validate if the application of the new model coefficient values, adapted to actual formats, brings a real progress in terms of correlation with subjective scores.

Codec	Resolution	Bit rate @ framerate
H.264 Baseline Profile	VGA	64@15fps
H.264 High Profile	(640 × 480)	128@15fps
		256@15fps
		384@15fps
		576@15fps
		128@30fps
		256@30fps
		384@30fps
		576@30fps
		768@30fps

Table 5.20: Database conditions

Correlation	G.1070	Extended G.1070
Pearson	0.82	0.92
Spearman	0.67	0.85
RMSE	0.81	0.64

Table 5.21: Correlations between G.1070 and subjective scores

We collected a database of audiovisual sequences elaborated during a subjective test on videotelephony scenarios carried out by Orange. The conditions of this test

correspond to some of the use cases concerned by this extension and are presented in Table 5.20.

We applied the G.1070 model as described on the ITU recommendation and the extension version on our database. Then we calculated the correlation coefficients between the two G.1070 scores and the subjective scores (see Table 5.21).

By comparing the correlation results, we note that extending the model with specific coefficients for the H.264 coding in Baseline profile and High profile conditions yields video quality scores closer to the subjective scores.

Chapter 6

Machine Learning approach for global no-reference video model generation

Contents

Introduction

The evaluation of video quality is a complex task given the multiplicity of parameters impacting the perceived media. The quality assessment subjective tests methodology, despite giving the exact perception of the quality, could not be used in real time. On the other hand, we have shown in Chapter 5 that the objective tools and models are numerous and that there is no representative metric for all degradation conditions.

In our study context of videoconferencing and video telephony services, we have shown through our subjective tests in Chapter 3 that the global audiovisual quality is generally more influenced by the video quality than the audio quality. This is why we focus mostly on assessing video quality of a videoconferencing service in real time. In this case, we consider no-reference metrics studied in Chapter 5 since in real time application reference signal is not available. Each of these metrics allow to measure the level of a single type of distortion impacting a video signal. However, the human perception of the quality does not distinguish between the types of distortion but it gives a global appreciation of the quality. Our idea is then to try

to combine all the MOAVI single artifact based metrics into a global video quality model generated by Machine Learning (ML) methods.

Machine Learning (ML) consists in the design and development of programs and algorithms which have the capability to automatically improve their performance on the basis of either their own experience over time, or earlier data provided by other programs [187]. We distinguish two types of Machine Learning algorithms: unsupervised and supervised learning. The unsupervised algorithm consists in estimating the structure of an unlabeled data. The use case of an unsupervised algorithm is the classification of data into categories. On the other side, the supervised learning is used when the category structure of the database is already known. Thus, the supervised learning predicts a function or a model that maps the database to the predefined class labels. In our case we are considering supervised learning, and we are interested in classification methods because of the discrete and labeled nature of our dataset and because our objective it to predict a variable.

In this chapter we present Machine Learning techniques for modeling the dependencies of different video impairments to the global video quality perception using subjective quality feedback.

6.1 Data mining tool

For our machine learning and data mining studies we used a software called "Khiops" [188]. The Khiops tool integrates the work done at Orange Labs on data preparation, automatic variable construction for multi-table databases and large-scale modeling.

Khiops allows to quickly perform the descriptive and explanatory phases in a Data Mining project. The database must be formatted according to a text file format, with a line per record, one header line containing the variable names and a field separator (tabulation by default).

The first step is the specification of the data dictionary, which is the choice of the variable types (Categorical, Numerical, Date, Time or Time stamp) in the database to analyze. This dictionary is automatically built by Khiops owing to a parsing of the database file. The built dictionary is saved in a dictionary file, which basic syntax allows easy modifications. The Data Miner must then validate the variable types in the built dictionary, and eventually specify which variables to ignore in the analysis or construct new variables owing the derivation rule language.

The second step checks the correctness of the database file. In this step, Khiops parses the database file and completely checks formatting or variable type errors.

The third step, the most important one, is to analyze the predictive value of the

explanatory variables or pairs of variables. In supervised analysis, when a target variable is specified, Khiops evaluates the predictive importance of any numerical or categorical explanatory variable, and of any pair of explanatory variables. Two reports, for uni-variate and bi-variate analysis, are produced at the end of the data analysis, based on the train data set. They summarize the information contained in each analyzed variable or pair of variables. In the case of supervised tasks, a scoring model is computed as well, based on a Selective Naive Bayes predictor. A modeling report summarizes the features of the built classifier or regressor. Two evaluation reports, based on the train and test data, evaluate the performance of the scoring model. New dictionaries and scoring dictionary, are produced, allowing a deployment of the scoring model.

The fourth step is the deployment step. This is done by applying the new dictionary or the scoring dictionary on new data, in order to compute score variables. This functionality can also be used to construct any new variable, described using the derivation rule language.

6.2 Descriptive analysis

The performance of each model generated by machine learning method depends directly on the used training database. The more database contains values representative of the final use cases and conditions, the more accurate the predictive model is. Thus, in our case we collected all the subjective databases available to us (either from our subjective tests or public databases) and presented in Chapter 3. Our training database consists in a total of 1130 data lines. Each line consists in a video sequence on which all the MOAVI metrics are applied and a subjective MOS score is associated.

6.2.1 Target variable

For our model, the variable that we try to predict is the subjective MOS score that we consider as the "Target variable". Since the MOS is a numerical and continuous variable, it must be discritized in order to be considered by the ML algorithm. Thus, from MOS values we associate new variable that we call "Quality" having four values:

- Excellent: if $MOS \geq 4$
- Good: if $3 \leq MOS \geq 4$
- Fair: if $2 \leq MOS \geq 3$
- Bad: if $MOS \leq 2$

We fixed this division because it is the one that gives the best balanced values distribution as shown in Figure 6.1

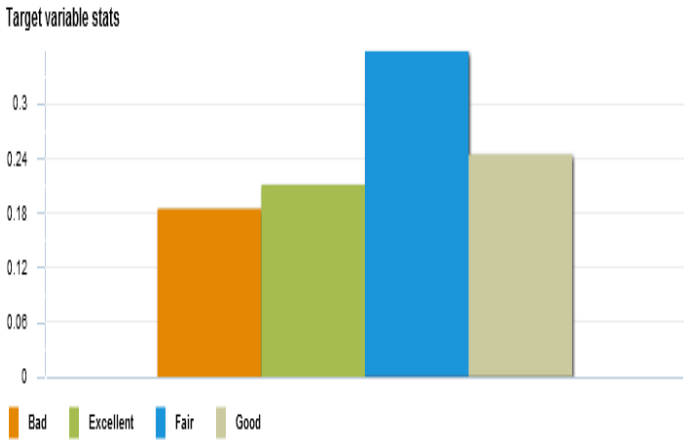


Figure 6.1: Target variable distribution

6.2.2 Outliers treatments

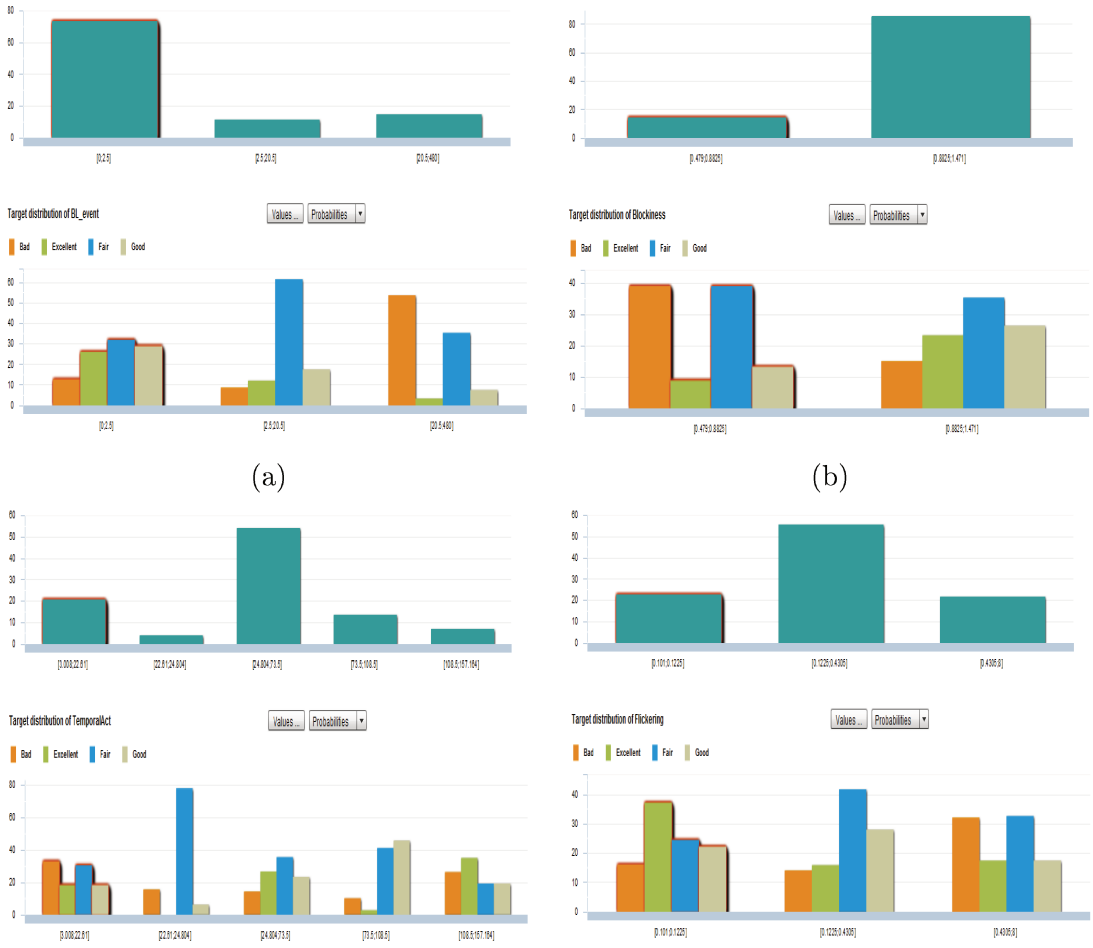
We consider a value as outlier when it stand out too much from the values generally observed on a variable. The process carried out concerned all the variables to remove the clearly incoherent values that are outside the range of 95% of the confidence interval. Visualization of the distribution and of the evolution of the mean of the variables, made it possible to judge the relevance of keeping or not these values in the sample. This treatment removed outliers that accounted for only less than 1% of the sample.

Before processing to the training of the ML model it is essential to have an idea on the distribution of the variables according to the Target as shown in Figure 6.2 (in these representations we draw our attention on the fact that the presentation of the four levels of the target variable does not follow the order of evolution of the quality.)

6.3 Selective naive Bayes model: obtaining a global video quality score

In our case, the variable to predict is the "Quality" metric defined above. Given the categorical nature of this target variable, we have the choice between a number of ML prediction methods, such as decision trees, random forests, and so on. Our choice is the Selective Naive Bayes (SNB) method because of:

- its simplicity,
- it is adapted to large volumes of data,
- its good performance often rented in publications [189],
- it is implemented in the software (Khiops) that we used.



(a) (b)

(c) (d)

Figure 6.2: Target variable distribution

6.3.1 Model results

As input for the ML algorithm we consider all the MOAVI metrics without doing a pre-selection of only the most correlated ones with subjective scores, as found in Chapter 5. However, the SNB algorithm defines the variables that are the most related to the MOS scores through an indicator called "Level". The level represents the evaluation of the predictive importance of the variable. It is a value between 0 (variable without predictive interest) and 1 (variable with optimal predictive importance). Figure 6.3 shows the distribution of the level values of our variables. The most correlated variable with subjective scores in our database is clearly Block loss event.

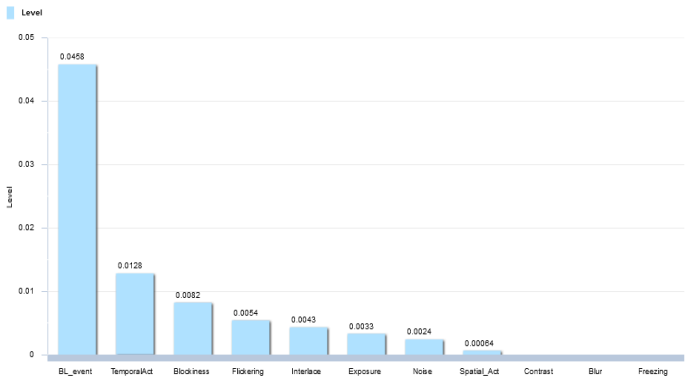


Figure 6.3: Level distribution

In Khiops datamining tool we fixed 70% of the database used for training and 30% for testing. The samples are chosen randomly by the algorithm. The table presented in Figure 6.1 shows the predictor evaluation on the test and training samples. The SNB classifier is evaluated using the following criteria:

- Accuracy: evaluates the proportion of correct prediction.
- Compression: evaluates the predicted target probabilities using a negative log likelihood approach and is normalized (between 0 and 1) using the baseline predictor.
- AUC: area under the ROC curve (AUC) which evaluates the ordering of the predicted scores per target value.

For our generated model we have 0.44 of accuracy, 0.09 for compression and 0.69 AUC which corresponds not to a fine prediction. According to these evaluation indicators, the generated model is not accurate for video quality assessment.

A confusion matrix is reported for the classifier, to compare the predicted values (prefixed by $) and the actual values ones. As shown in Figure 6.4, for Bad and Good values, the model gives a correct prediction in 70% of the cases. However, for Fair and Excellent the model gives a correct prediction in less than 50% of the cases. This can be explained by the fact that Fair and Good classes are close to

Name	Type	AUC	Compression	Accuracy
Selective Naive Bayes	Train	0.7077	0.1341	0.5082
Optimal	Train	1	1	1
Selective Naive Bayes	Test	0.6915	0.0917	0.4438
Optimal	Test	1	1	1

Table 6.1: Predictor evaluation

target	%Bad	%Excellent	%Fair	%Good
$Bad	70.00	0.00	17.50	12.50
$Excellent	14.67	42.39	21.74	21.20
$Fair	13.76	17.03	48.03	21.18
$Good	0.00	16.42	13.43	70.15

Figure 6.4: Confusion matrix

each other and over-represented in our database.

Moreover, the cumulative gain curve, drawn in Figure 6.5, evaluates the quality of the model. The green curve corresponds to the results of the SNB model applied on the test sample database. The purple one corresponds to an optimal model. The black curve corresponds to the worst model, that is to say the one that is equivalent to a random choice of the class.

Based only on MOAVI single artifact based metrics it is shown that the ML approach generate a model that is not accurate in predicting the global video quality. Thus, we have the idea to add another no-reference metric to the training variables which is VIIDEO. This metric will bring information on the global quality of the sequence (not dedicated for a specific distortion) that could enhance the performance of the prediction model.

We apply the same methodology as described above, we add the VIIDEO metric values for all our 1130 sequences and we re-run the training and testing processes. The evaluation of the generated prediction model presented in Table 6.2 shows clearly that the accuracy of the model is improved 0.618.

The new confusion matrix presented in Figure 6.6 shows that the new model makes less error in quality prediction compared to the one trained only on MOAVI metrics. For Fair and Excellent classes the model gives more that 50% of correct prediction. For the Good class it gives 79.31% of correct prediction. For the Bad class it reaches 91% of correct prediction which is particularly interesting because for a monitoring and diagnostic tool it is important to detect a Bad quality when

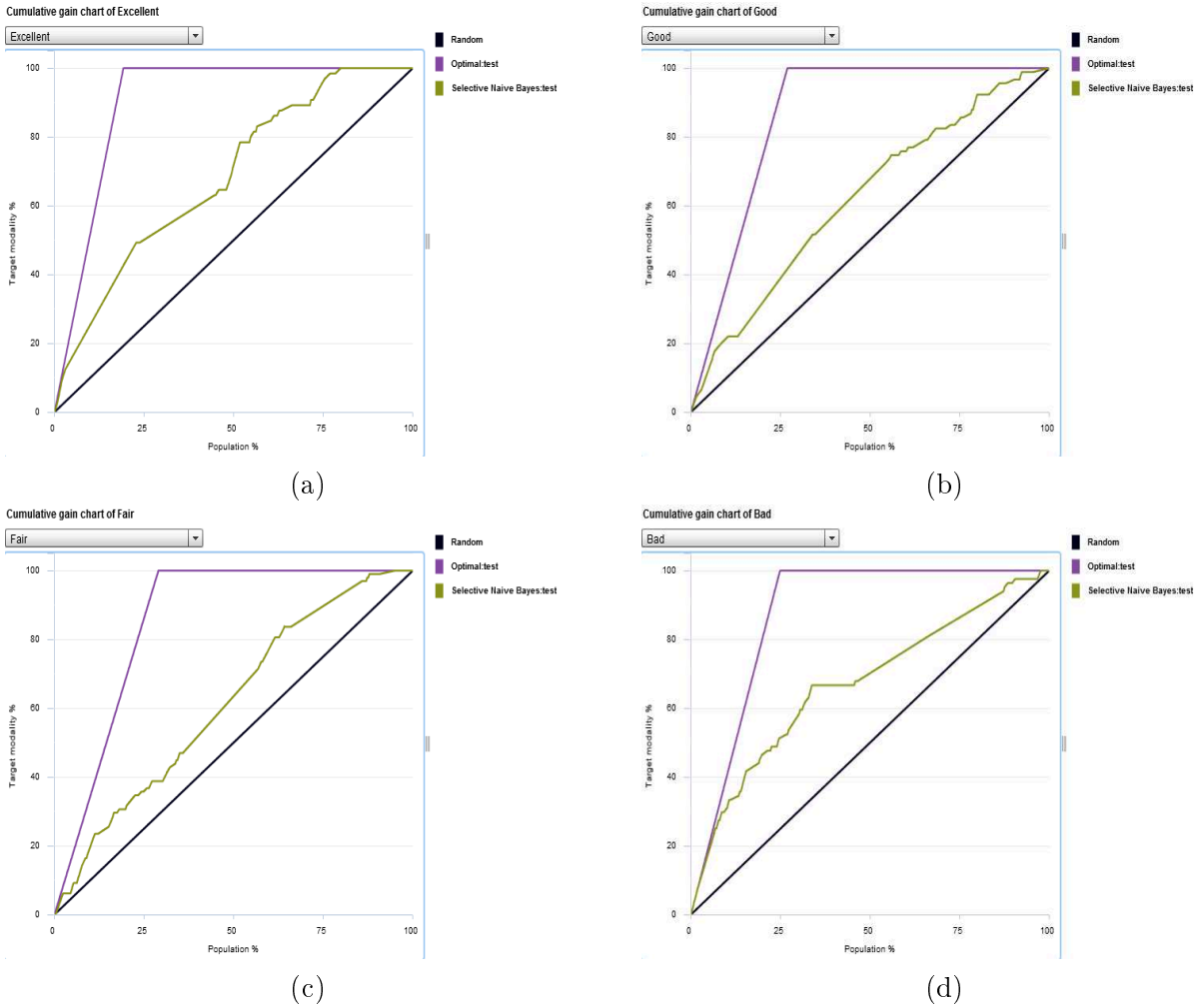


Figure 6.5: Cumulative gain curve for Excellent (a), Good (b), Fair (c) and Bad (d) quality classes

Name	Type	AUC	Compression	Accuracy
Selective Naive Bayes	Train	0.8038	0.3739	0.6350
Optimal	Train	1	1	1
Selective Naive Bayes	Test	0.8097	0.3597	0.618
Optimal	Test	1	1	1

Table 6.2: Predictor evaluation after adding VIIDEO metric

target	%Bad	%Excellent	%Fair	%Good
$Good	0.00	6.90	13.79	79.31
$Fair	0.48	18.90	54.78	25.84
$Excellent	0.00	52.56	23.72	23.72
$Bad	91.72	1.27	5.73	1.27

Figure 6.6: Confusion matrix

there is a problem more than to detect a good quality.

The good performance of the model is confirmed by the cumulative gain curves shown in Figure 6.7. The green curves corresponding to the SNB predictor are close to the optimal predictor.

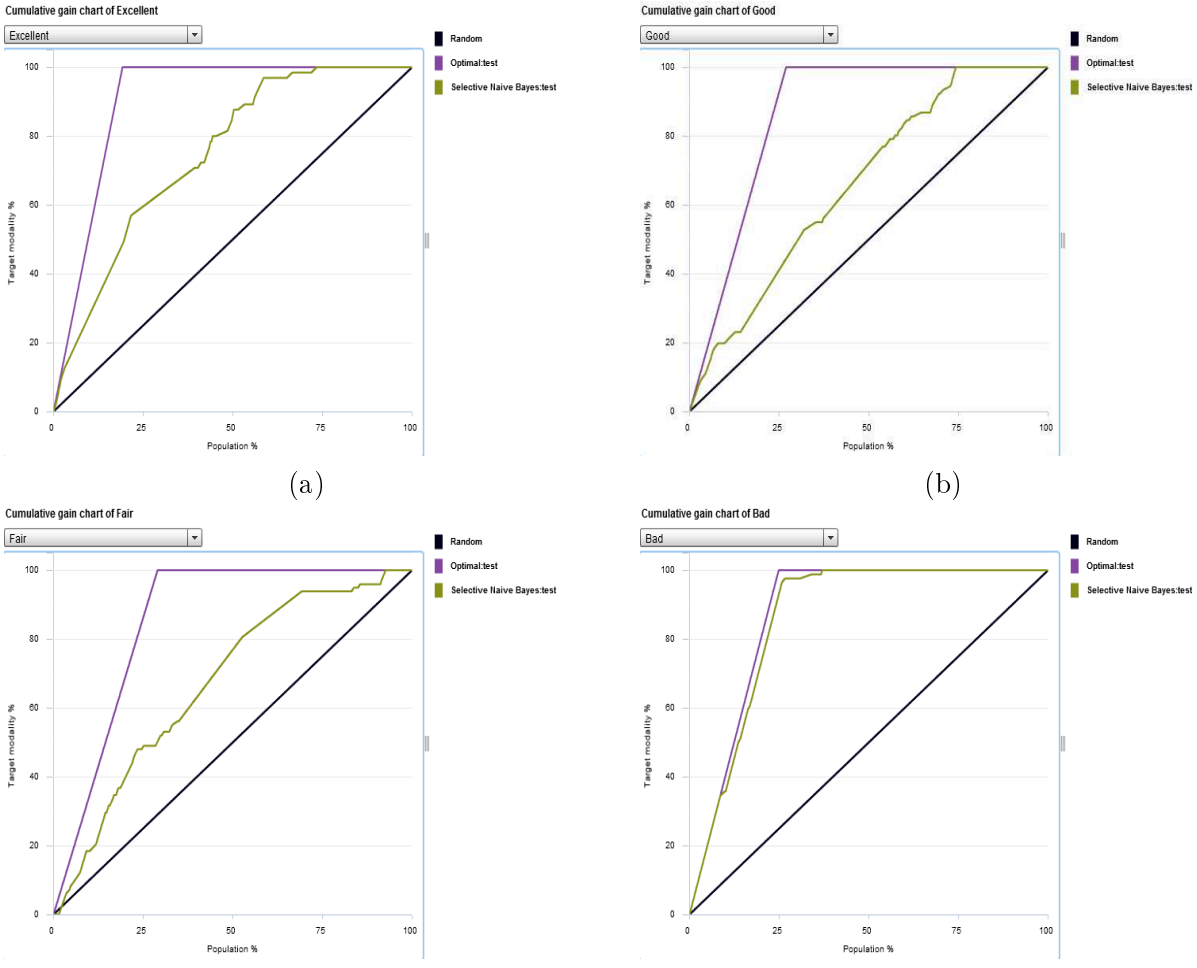


Figure 6.7: Cumulative gain curve for Excellent (a), Good (b), Fair (c) and Bad (d) quality classes

6.4 Conclusion

In this chapter we investigate the possibility of combining no-reference single artifact metrics taken from MOAVI in a global video quality assessment model. The obtained model has an accuracy of only 0.44 which is not enough for a good model. After adding no reference VIIDEO metric to the training variables of the ML algorithm, the model is enhanced and reach 0.63 of accuracy. This result is encouraging because we consider that even if our database contains only 1130 sequences, this volume allowed to generate a promising prediction model. We recommend to collect more databases with more diversified conditions.

Conclusion and perspectives

The work carried out in this thesis led to several results in the field of QoE in the context of video telephony and videoconferencing services. The contributions are twofold, as they relate to both subjective and objective evaluation of the audiovisual quality of a video call. The first contribution is the constitution of a database of audio visual sequences corresponding to a real scenario of video call, a crucial question for the audiovisual quality community. The second contribution concerns the evaluation of the existing objective quality assessment tools.

Contributions to the subjective assessment of the audiovisual quality

Firstly, we carried out two modalities of subjective audiovisual quality test. Our objective was investigating audiovisual quality in both interactive and non-interactive contexts and under different scene complexities. By comparing non-interactive vs. interactive test results, we found that statistically there is no significant difference for MOS_A, MOS_V and MOS_{synch} scores between the two experimental contexts. Nevertheless, considering MOS_{AV} scores we noted a significant difference between the two contexts. Thus, in future experiments we can rely on non-interactive test only and apply their results (with the exception of the evaluation of AV quality, for which interactive tests remain mandatory) to a conversational context. Besides, the results show that the scene complexity has an impact on the perceived audiovisual quality in both contexts and on the perception of audio-video synchronization in the interactive context.

Secondly, we were interested in better understanding the influence of the time offset between the audio and the video media streams of videotelephony contents in or without the presence of other impairments (packet loss, encoding, resolution change). Our tests allow us to define the acceptability thresholds of audio/video delay for video telephony context under various conditions. Our subjective tests show that the spatial complexity of video contents and noisiness of audio contexts have a negative influence on the perception of desynchronization (the more complex or the more noisy, the worse in terms of perception). Furthermore, in presence of loss of audio and video information (resulting from IP packet loss), the desynchronization is less perceptible.

Thirdly, a comparison between laboratory and crowdsourcing subjective scores results show that the use of a crowdsourcing approach leads to results without significant difference compared with those of the laboratory test for the assessment of the video quality and global audiovisual quality on various types of contents representative of a videotelephony conversation. Nevertheless, the results are less promising as far as the desynchronization perceptibility and (mostly) the audio quality are

concerned, where the crowdsourcing approach yields underestimation of the quality and lower discrimination between bad and good conditions. The difference between the media rendering hardware used in laboratory and at home, as well as the uncontrolled acoustic environment, are certainly the main reasons.

Contributions to the objective assessment of the audiovisual quality

In this thesis we have conducted an updated survey of the developed media-layer full reference objective video quality models. We carried out a performance comparison of ten different objective metrics in the context of video calling and videoconferencing. Experimental results show that metrics which include information about temporal video aspect in the quality estimation algorithm outperform other metrics. For the databases containing only packet loss transmission errors, all the metrics are well correlated with the subjective video quality perception, with a little preference for OPVQ, Vis3, SSIMplus and VMAF. For databases including contents closer to those in real videoconferencing context, ViS3 statistically outperforms the other tested metrics.

In what concerns impairments caused by video encoding bitrates, VMAF and SSIMplus are the most competitive metrics. When considering all degradation types across all databases, OPVQ, VMAF, ViS3 and SSIMplus have an equal statistical performance that exceeds the other metrics. For evaluating the influence of codec type, coding bitrate and frame rate changes, OPVQ, ViS3, SSIMplus and VMAF may give out objective scores better correlated with the MOS. However, further studies are needed to optimize the OPVQ algorithm for the new generation of video codecs such as the HEVC. In the case of network transmission errors, we have a high probability to obtain a temporal misalignment between the reference and the degraded sequences. As a result, the scores of metrics based on frame by frame comparison are biased. In that case, we recommend the use of the ViS3 metric because its algorithm is based on computing quality on the GOP and the STS. VMAF is a promising model for video quality since it is constructed using the machine learning approach. Its performance can be ameliorated by enriching the learning data set with large simple of impairment and contents types, and by training other better objective metrics such as the SSIMplus, ViS3...etc. Our experimental results show that there is no universal metric which is best for all distortion types and contents.

In the context of real time video quality assessment, no reference metrics are recommended. Thus, we evaluated the performance of six no-reference single artifact based video quality assessment metrics developed by MOAVI VQEG project. We find that the metrics may be representative indicators of video quality. For each condition (encoding, packet loss, signal attenuation, etc) we identified the representative metrics that we recommend to take into account. According to the obtained results, it can be seen that for transmission impairments, distortions perceived by end user can be well reflected by metrics representative of block loss events, slicing

or freezing. In what concerns impairments related to encoding, they are essentially linked with metrics on blur, blockiness and flickering. We think that these metrics, and associated thresholds, constitute an unavoidable part of the tool box to diagnose video quality in communication services.

During the last months of this thesis we proposed a methodology for modeling video quality using Machine Learning approach. We investigated the possibility of combining no-reference single artifact metrics in a global video quality assessment model. The obtained model has an accuracy of only 0.44 which is not enough for a good model. After adding no reference VIIDEO metric to the training variables of the ML algorithm, the model is enhanced and reach 0.63 of accuracy. This result is encouraging because we consider that even if our database contains only 1130 sequences, this volume allowed to generate a promising prediction model. We recommend to collect more databases with more diversified conditions.

Perspectives

In general, our work has led to a better understanding of audiovisual quality assessment processes for videotelephony services. Nevertheless, there are still some grey areas to clear up and the possibility of deepening some of the proposed approaches. Most important, broadening the base of audiovisual sequences would allow better learning of objective criteria. This would also reduce the inaccuracies on the performance indicators.

In all our subjective tests, we have limited ourselves to the evaluation of application type and transmission impairments. It is obvious that a video telephony service is impacted by other factors, such as context, psychological situation, type of terminal, OS .... Enlargement to a wider spectrum of impairments and conditions would allow a finer characterization of the quality of a video call service.

Several additional works are feasible on the objective quality criteria, in particular in the development of real-time solutions. We believe that the Machine Learning approach is promising. It is possible to collect a larger video quality training database in order to cover all the possible degradations of video quality. Furthermore, we trained our model only on MOAVI no-reference metrics and VIIDEO metric. It would be interesting to investigate additional variables such as, video resolution, coding bit rate, percentage of packet loss, etc. These variables bring additional informations to the algorithm and make it more decisive.

Chapter 7

WebRTC architecture

In the WebRTC architecture model, both browsers are running a web application, which is downloaded from the same web server. Signaling messages are used to set up and terminate communications. They are transported by the HTTP or WebSocket protocol via web servers that can modify, translate, or manage them as needed. It is worth noting that the signaling between browser and server is not standardized in WebRTC, as it is considered to be part of the application (see Signaling). As to the data path, a PeerConnection allows media to flow directly between browsers without any intervening servers. The two web servers can communicate using a standard signaling protocol such as SIP or Jingle (XEP-0166). Otherwise, they can use a proprietary signaling protocol.

A WebRTC web application (typically written as a mix of HTML and JavaScript) interacts with web browsers through the standardized WebRTC API, allowing it to properly exploit and control the real-time browser function. The WebRTC API must therefore provide a wide set of functions, like connection management (in a peer-to-peer fashion), encoding/decoding capabilities negotiation, selection and control, media control, firewall and NAT element traversal, etc.

The API is being designed around three main concepts: MediaStream, PeerConnection, and DataChannel.

- **MediaStream**: is an abstract representation of an actual stream of data of audio and/or video. It serves as a handle for managing actions on the media stream, such as displaying the stream's content, recording it, or sending it to

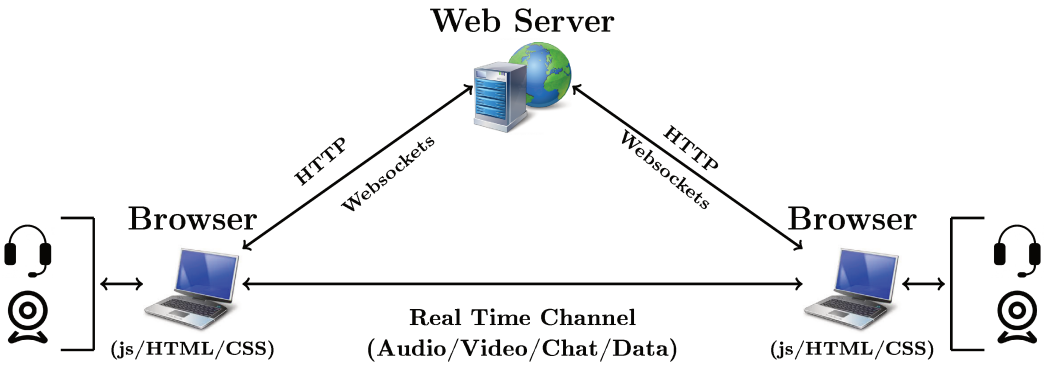


Figure 7.1: WebRTC triangle architecture

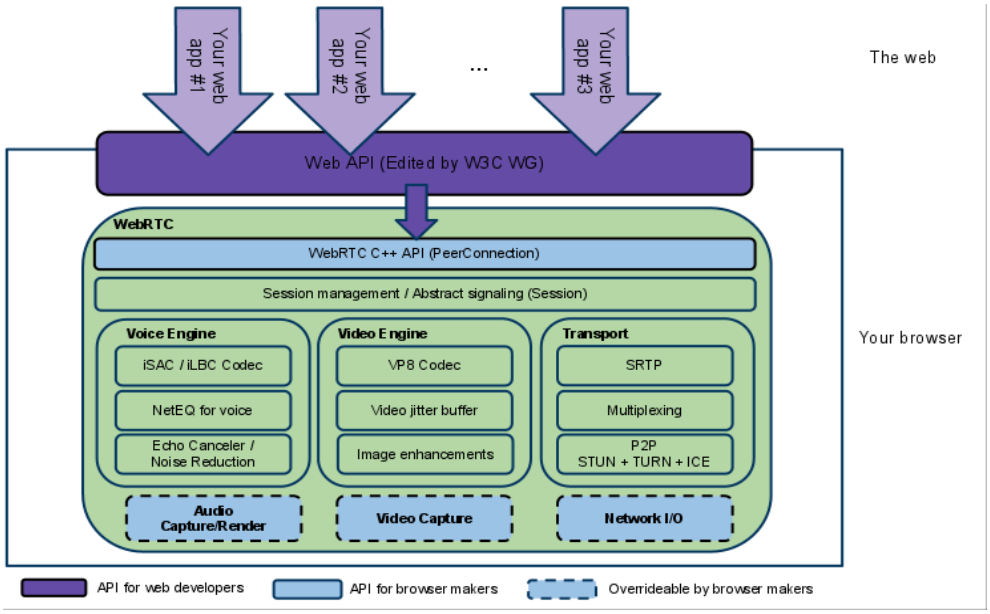


Figure 7.2: WebRTC architecture in the browser level (source:https://webrtc.org/architecture/

a remote peer. A MediaStream may be extended to represent a stream that either comes from (remote stream) or is sent to (local stream) a remote node.

- **PeerConnection**: allows two users to communicate directly, browser to browser. It then represents an association with a remote peer, which is usually another instance of the same JavaScript application running at the remote end. Communications are coordinated via a signaling channel provided by scripting code in the page via the web server, e.g., using XMLHttpRequest or WebSocket. Once a peer connection is established, media streams (locally associated with ad hoc defined MediaStream objects) can be sent directly to the remote browser.

- **DataChannel**: The DataChannel API is designed to provide a generic transport service allowing web browsers to exchange generic data in a bidirectional peer-to-peer fashion.

Chapter 8

Libon database QoE analysis

We have collected from Orange a database constituted of 19000 Libon calls established on a period of 1 month in 2015. The database is constituted from technical indicators collected from the terminal and the network and subjective scores given by the client evaluation of the quality by the end of the call. Figure 8.1 shows the Libon pop-up application that allows collecting users evaluation. All the database quality indicators are summarized in Table 8.1.

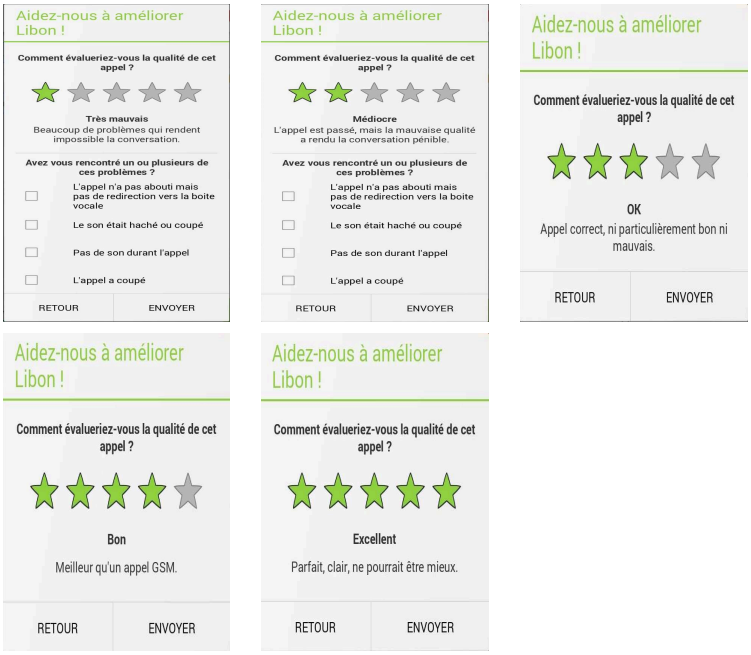


Figure 8.1: Screenshot from Libon questionnaire for quality evaluation

With the available database we investigated statistical correlations between the indicators. Thus, we started with a global analysis where our objective is to determine a limited list of metrics with strong correlation with the perceived quality by user, and consider this short list for the advanced studies.

Global analysis

Target.
"Rating score".

Field	Description
Devise	Type of the user devise
OS	Type of the user Operation System OS : Android
OS version	Version of the used OS
Country	Country of the caller (retrieved only when GPS in activated)
Network type	Type of network used during the call: Wifi, 2G, 3G or 4G
Networkoperator	Name of network operator
Date	Date of the call
Time	Time of the call
Direction	Outgoing call or Incoming call
Duration	Call duration in seconds
State	End call state: aborted, declined, missed or success
calltype	Call type: VoipOut/App2app/CovExt/VoiceMail/Unknown
Codec	Used audio codec: iLB, opus or PCMA
EchoDesc	Echo canceler used(Builtin/NOAEC/WebRTCAEC/SpeexEC)
packetsrecv	Number of received packets
packetssent	Number of sent packets
avgestlatency	Average latency
maxcpu	Maximum of CPU usage
averagecpu	Average CPU usage
ratingscore	The number of stars the user gave to the call 1-5, if the call did not go through, or the user did not rate the call, this will be N/A.
ratingcomments	Comments written by the client for diverse remarks
RatingCallHadDelay	Indicates if the user experienced latency. If the user rated the call, this will be true or false, depending on if the checkbox was ticked
RatingCallTruncated-	Indicates if the sound was truncated or stuttering. If the user rated the call, this will be true or false, depending on if the checkbox was ticked
Stuttering RatingBack-groundNoise	Indicates if the user experienced background noise. If the user rated the call, this will be true or false, depending on if the check-box was ticked
RatingCallDropped	Indicates if the call was dropped. If the user rated the call, this will be true or false, depending on if the checkbox was ticked
moscq	libon conversational quality indicator
moslq	libon listening quality indicator
AvgMos	average bidirectional MOS also considering RTT
AvgMosRx	average MOS of the received stream
AvgMosTx	average MOS of the transmitted stream

Table 8.1: Some fields of Libon voice call report

Methodology.

Search for correlation with all other fields in the database.

A clustering of variables will be also performed, i.e. not only a study of iso-

lated data. This global analysis is completed with an analysis, for each interesting variable, of value distributions.

Results.

First of all, the distribution of the subjective note "Rating score" considered as the target variable is presented in Table 8.2.

Rating score	Percentage (%)
5	53.70%
4	13.97%
3	13.34%
2	7.51%
1	11.49%

Table 8.2: Distribution of the Rating score

We note that the given score 5 is dominant and we found this aspect even if we make restricted studies on specific profiles. In order to verify if the dominance of the value 5 reflects really a good quality of calls or the users asked to the evaluation question randomly, we studied the distribution of the objective parameters "moscq" and "moslq". The distribution of the variable Mos_CQ is as follows:

MOS_CQ	Percentage (%)
[4.25, 5]	65.24%
[3.95, 4.25]	12.48%
[3.35, 3.95]	11.59%
[1, 3.35]	10.68%

Table 8.3: Distribution of the MOS_CQ

A direct correlation of "Rating score" with all other fields in the tickets gives no obvious correlation. The best level of correlation is found, without surprise, with:

- other subjective scores (open for answer only if "Rating score" is below 3),
- moscq, packetsrecv, moslq, mcc

The current indicators used by Orange which are the Average Call Duration (ACD) and MOS-CQ do not show any particular level of correlation with the rating score. Taking all the metrics and calculating the correlations between them and the RatingScore we obtained low levels of correlations. The first five most correlated variables are: caller, moscq, packetsrecv, mcc, moslq.

Thus, the analysis must be focused on more representative data. We followed two possible tracks in order to have a specific data sample:

- exclusion of all rating scores above 4 (no significant improvement),
- restriction to some interesting profiles, based on the following metrics

- Devices: selection of the TOP 20 devices, 19 of them being Samsung Galaxy models.
- Call types: abandoned, since more than 90 % of calls are out-going VoIP calls.
- Network: WiFi is widely represented (60.58%), 3G (21.86%), 4G (11.17%) and only a few samples for 2G.
- Countries: possibility of regional clustering (France, Western Europe, North America, Arabic peninsula). 38.13% of calls are made from Europe (France in the first place and then Spain and Italy) then Asia 21.64%, America 6% and only 3.91% from Africa. The important distribution of the data and the few amount of samples decreases the correlation levels.

Analysis on other subjective data as a target

Targets of consideration:

- RatingCallHadDelay
- RatingCallTruncatedStutering
- RatingBackgroundNoise
- RatingCalldropped

Results.
We obtained no enough data. In fact, these questions are asked only when the "Rating score" is below 3. In reality this means a very small amount of data: 424 users encountering delay, between 200 and 300 users for other degradations and only 33 written comments.

Analysis on objective MOS as a target

Here we take the assumption that MOSCQ is a good and representative objective indicator on the quality of experience. We consider it as the target for the analysis. Results show that the best correlation between MOS_CQ and all other fields is found with the other MOS values : Mos_LQ, avgmos, avgmosrx ... . Mos_CQ is more correlated with the technical parameters (jitter, latency, packet loss ...) than the RatingScore. On the other side there is no correlation with the context parameters (device, os, caller, contry ...)

Chapter 9

Subjective audiovisual test questions

Questions asked in the audiovisual subjective tests:

1. How do you rate the global audiovisual quality ?
   - Bad
   - Poor
   - Fair
   - Good
   - Excellent
2. How do you rate the video quality ?
   - Bad
   - Poor
   - Fair
   - Good
   - Excellent
3. How do you rate the audio quality ?
   - Bad
   - Poor
   - Fair
   - Good
   - Excellent
4. How do you rate desynchronization between the image and the sound ?
   - Very annoying
   - Annoying
   - Slightly annoying
   - Perceptible but not annoying
   - Imperceptible

Figure 9.1 below shows the exact labels (in French) used for these tests.

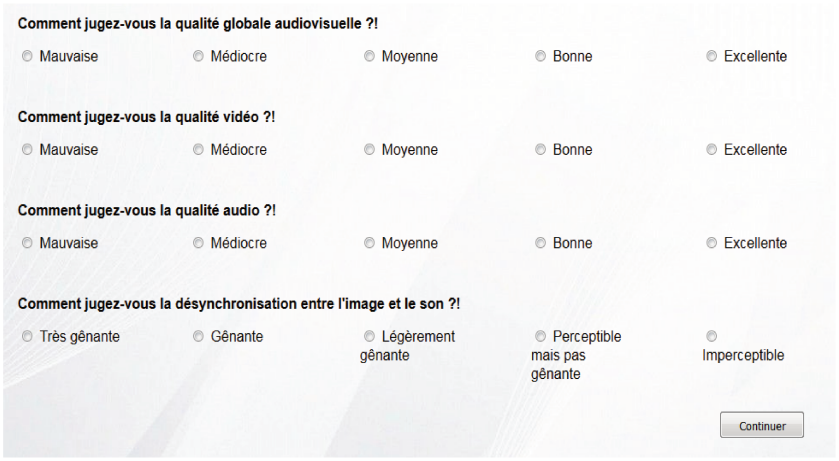


Figure 9.1: labels of questions (in French)

Bibliography

[1] Margaret H Pinson and Stephen Wolf, "A new standardized method for objectively measuring video quality," *IEEE Transactions on broadcasting*, vol. 50, no. 3, pp. 312–322, 2004.

[2] "No reference metrics," http://vq.kt.agh.edu.pl/metrics.html.

[3] Cisco Visual Networking Index Cisco, "Global mobile data traffic forecast update, 2015–2020 white paper, 2016," .

[4] "Skype usrs," https://mspoweruser.com/skype-300-million-monthly-active-users/.

[5] "Skype calls," https://news.microsoft.com/bythenumbers/skype-calls.

[6] "Utra-utran long term evolution (lte) and 3gpp system architecture evolution (sae)," ftp://ftp.3gpp.org/Inbox/2008_web_files/LTA_Paper.pdf, 2008.

[7] "3gpp, the third generation partnership projec," http://www.3gpp.org/.

[8] GSM Association IR 94, "Ims profile for conversational video service," 2013.

[9] speech processing transmission ETSI Guide 202 057-2 and quality aspects (stq), "User related qos parameter definitions and measurements; part 2: voice telephony, group 3 fax, modem data services and sms," Tech. Rep., Technical report, ETSI, 2009.

[10] 3GPP TS 23.203, "Policy and charging control architecture," 2014.

[11] "Décision n 2013 - 0004 de l'autorité de régulation des communications électroniques et des postes en date du 29 janvier 2013 relative à la mesure et à la publication d'indicateurs de la qualité d es service s fixes d'accès à l'in," https://www.arcep.fr/uploads/tx_gsavis/13-0004.pdf.

[12] Report of the Regulatory Authority for Electronic Communications and Posts, "La qualité des services mobiles en france métropolitaine. les résultats de l'enquête 2014," http://www.arcep.fr/uploads/tx_gspublication/rapport-QS-mobile-2014-230614.pdf.

[13] "Itu-t," http://www.itu.int/en/ITU-T/publications/Pages/recs.aspx.

[14] ITU-R Rec. P.863, "Perceptual objective listening quality assessment," 2014.

[15] "Web site polqa presentation," http://www.polqa.info/.

[16] "Webrtc," https://webrtc.org/.

[17] "Webrtc 1.0: Real-time communication between browsers," https://www.w3.org/TR/webrtc/.

[18] "Browser support scorecard," http://iswebrtcreadyyet.com/.

[19] "Javascript session establishment protocol," https://tools.ietf.org/html/draft-ietf-rtcweb-jsep-18.

[20] "Orange libon service," https://www.libon.com.

[21] Kalevi Kilkki, "Quality of experience in communications ecosystem.," *J. UCS*, vol. 14, no. 5, pp. 615–624, 2008.

[22] Sebastian Möller and Alexander Raake, *Quality of experience: advanced concepts, applications and methods*, Springer, 2014.

[23] Kjell Brunnström, Sergio Ariel Beker, Katrien De Moor, Ann Dooms, Sebastian Egger, Marie-Neige Garcia, Tobias Hossfeld, Satu Jumisko-Pyykkö, Christian Keimel, Mohamed-Chaker Larabi, et al., "Qualinet white paper on definitions of quality of experience," 2013.

[24] Lawrence A Pervin and Oliver P John, *Handbook of personality: Theory and research*, Elsevier, 1999.

[25] W. Robitza and A. Raake, "(re-)actions speak louder than words? a novel test method for tracking user behavior in web video services," in *Proc. QoMEX*, June 2016.

[26] Mojca Volk, Janez Sterle, Urban Sedlar, and Andrej Kos, "An approach to modeling and control of qoe in next generation networks [next generation telco it architectures]," *IEEE Communications Magazine*, vol. 48, no. 8, 2010.

[27] Peter Reichl, Bruno Tuffin, and Raimund Schatz, "Logarithmic laws in service quality perception: where microeconomics meets psychophysics and quality of experience," *Telecommunication Systems*, pp. 1–14, 2013.

[28] Michal Ries, Peter Froehlich, and Raimund Schatz, "Qoe evaluation of high-definition iptv services," in *RADIOELEKTRONIKA*, 2011, pp. 1–5.

[29] Sabina Baraković, Jasmina Baraković, and Himzo Bajrić, "Qoe dimensions and qoe measurement of ngn services," in *Proc. TELFOR*, 2010.

[30] Satu Jumisko-Pyykkö, "User-centered quality of experience and its evaluation methods for mobile television," *Tampere University of Technology*, p. 12, 2011.

[31] Ulrich Reiter, Kjell Brunnström, Katrien De Moor, Mohamed-Chaker Larabi, Manuela Pereira, Antonio Pinheiro, Junyong You, and Andrej Zgank, *Factors Influencing Quality of Experience*, pp. 55–72, Springer International Publishing, Cham, 2014.

[32] Marwin Schmitt, Simon Gunkel, Pablo Cesar, and Peter Hughes, "A qoe testbed for socially-aware video-mediated group communication," in *Proc. ACM*, 2013, pp. 37–42.

[33] Markus Fiedler, Tobias Hossfeld, and Phuoc Tran-Gia, "A generic quantitative relationship between quality of experience and quality of service," *IEEE Network*, vol. 24, no. 2, 2010.

[34] James Nightingale, Qi Wang, Christos Grecos, and Sergio Goma, "The impact of network impairment on quality of experience (qoe) in h. 265/hevc video streaming," *IEEE Transactions on Consumer Electronics*, vol. 60, no. 2, pp. 242–250, 2014.

[35] Pradip Paudyal, Federica Battisti, and Marco Carli, "Impact of video content and transmission impairments on quality of experience," *Multimedia Tools and Applications*, vol. 75, no. 23, pp. 16461–16485, 2016.

[36] Mu Mu, Roswitha Gostner, Andreas Mauthe, Gareth Tyson, and Francisco Garcia, "Visibility of individual packet loss on h. 264 encoded video stream: a user study on the impact of packet loss on perceived video quality," SPIE, 2009.

[37] Thomas Zinner, Oliver Hohlfeld, Osama Abboud, and Tobias Hoßfeld, "Impact of frame rate and resolution on objective qoe metrics," in *Proc. QoMEX*, 2010, pp. 29–34.

[38] Demin Wang, Filippo Speranza, Andre Vincent, Taali Martin, and Phil Blanchfield, "Toward optimal rate control: a study of the impact of spatial resolution, frame rate, and quantization on subjective video quality and bit rate," in *Proc. of SPIE Vol*, 2003, vol. 5150, p. 199.

[39] David Geerts, Katrien De Moor, Istvan Ketyko, An Jacobs, Jan Van den Bergh, Wout Joseph, Luc Martens, and Lieven De Marez, "Linking an integrated framework with appropriate methods for measuring qoe," in *Proc. QoMEX*, 2010, pp. 158–163.

[40] Ina Wechsung, Klaus-Peter Engelbrecht, Christine Kühnel, Sebastian Möller, and Benjamin Weiss, "Measuring the quality of service and quality of experience of multimodal human–machine interaction," *Journal on Multimodal User Interfaces*, vol. 6, no. 1-2, pp. 73–85, 2012.

[41] Miguel Ríos Quintero and Alexander Raake, "Towards assigning value to multimedia qoe," in *Proc. QoMEX*, 2011, pp. 1–6.

[42] Khalil Ur Rehman Laghari and Kay Connelly, "Toward total quality of experience: A qoe model in a communication ecosystem," *IEEE Communications Magazine*, vol. 50, no. 4, 2012.

[43] Andreas Sackl, Kathrin Masuch, Sebastian Egger, and Raimund Schatz, "Wireless vs. wireline shootout: Howuser expectations influence quality of experience," in *Proc. QoMEX*, 2012, pp. 148–149.

[44] Md Abdur Rahman, Abdulmotaleb El Saddik, and Wail Gueaieb, "Augmenting context awareness by combining body sensor networks and social networks," *IEEE Transactions on Instrumentation and Measurement*, vol. 60, no. 2, pp. 345–353, 2011.

[45] S Rihs, "The influence of audio on perceived picture quality and subjective audio-video delay tolerance," *RACE MOSAIC deliverable*, , no. R2111, 1995.

[46] ITU-R Rec. BT.1359, "Relative timing of sound and vision for broadcasting," 1998.

[47] ITU-R Rec. SG11 11A/55, "Evaluation of the subjective effects of timing errors between sound and vision signals in television," 1995.

[48] ITU-T Rec. P.800, "Methods for subjective determination of transmission quality," August 1996.

[49] ITU-T Recommendation P.805, "Subjective evaluation of conversational quality," April 2007.

[50] ITU-T Rec. P.806, "A subjective quality test methodology using multiple rating scales," February 2014.

[51] ITU-T Rec. P.835, "Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm," November 2003.

[52] ITU-R Rec. BS.1284-1, "General methods for the subjective assessment of sound quality," 2003.

[53] ITU-R Rec. BS.1534-1, "Method for the subjective assessment of intermediate quality level of coding systems," 2003.

[54] ITU-R Rec. BS.1116-1, "Method for the subjective assessment of small impairments in audio systems including multichannel sound systems," 1997.

[55] ITU-R Rec. BT.500-13, "Methodology for the subjective assessment of the quality of television pictures," 2012.

[56] ITU-R Rec. BT.1788-0, "Methodology for the subjective assessment of video quality in multimedia applications," 2007.

[57] ITU-R Rec. BS.775-0, "Multichannel stereophonic sound system with and without accompanying picture," 2012.

[58] ITU-R Rec. BS.1286-0, "Methods for the subjective assessment of audio systems with accompanying picture," 1997.

[59] ITU-T Rec. P.910, "Subjective video quality assessment methods for multimedia applications," April 2008.

[60] ITU-T Rec. P.920, "Interactive test methods for audiovisual communications," May 2000.

[61] ITU-T Rec. P.911, "Subjective audiovisual quality assessment methods for multimedia applications," Dec. 1998.

[62] "International telecommunication union," http://www.itu.int/en/Pages/default.aspx.

[63] ITU-T Rec. G.1011, "Reference guide to quality of experience assessment methodologies," July 2016.

[64] Akira Takahashi, David Hands, and Vincent Barriac, "Standardization activities in the itu for a qoe assessment of iptv," *IEEE Communications Magazine*, vol. 46, no. 2, 2008.

[65] ITU-T Rec. J.247, "Objective perceptual multimedia video quality measurement in the presence of a full reference," Aug. 2008.

[66] ITU-T Rec. P.1201, "Parametric non-intrusive assessment of audiovisual media streaming quality," October 2012.

[67] ITU-T Rec. G.1070, "Opinion model for video-telephony applications," Aug. 2012.

[68] ITU-T Rec. G.1071, "Opinion model for network planning of video and audio streaming applications," Nov. 2016.

[69] ITU-T Rec. P.1202, "Parametric non-intrusive bitstream assessment of video media streaming quality," Oct. 2012.

[70] ITU-T Rec. P.1203, "Parametric bitstream-based quality assessment of progressive download and adaptive audiovisual streaming services over reliable transport," October 2017.

[71] ITU-T Rec. J.343, "Hybrid perceptual bitstream models for objective video quality measurments," Nov. 2014.

[72] Inès Saidi, Lu Zhang, Vincent Barric, and Olivier Déforges, "Evaluation of the performance of itu-t g. 1070 model for packet loss and desynchronization impairments," in *Proc. QoMEX*, 2016.

[73] Jose Joskowicz, J Carlos López Ardao, and Rafael Sotelo, "Quantitative modeling of the impact of video content in the ITU-T G. 1070 video quality estimation function," *Informática na educação: teoria & prática*, vol. 14, no. 2, 2011.

[74] Jose Joskowicz and J Ardao, "Enhancements to the opinion model for video-telephony applications," in *Proceedings of the 5th International Latin American Networking Conference*. ACM, 2009, pp. 87–94.

[75] Niranjan D Narvekar, Tao Liu, Dekun Zou, and Jeffrey A Bloom, "Extending G. 1070 for video quality monitoring," in *Proc. ICME*, 2011, pp. 1–4.

[76] Debajyoti Pal, Tuul Triyason, and Vajirasak Vanijja, "Extending the ITU-T G. 1070 opinion model to support current generation H.265/HEVC video codec," in *International Conference on Computational Science and Its Applications*. Springer, 2016, pp. 106–116.

[77] Debajyoti Pal and Vajirasak Vanijja, "G.1070 model extension at Full HD resolution for VP9/HEVC codec," *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, vol. 8, no. 9, pp. 139–147, 2016.

[78] "Itu contribution 129 sg12: Discussion of HP," .

[79] Huawei Technologies Co. Ltd., "Itu contribution 130 sg12: Discussion of BP training experiment results for G.1070 extension," .

[80] Quan Huynh-Thu and Mohammed Ghanbari, "Scope of validity of PSNR in image/video quality assessment," *Electronics letters*, vol. 44, no. 13, pp. 800–801, 2008.

[81] Z Wan and AC Bovik, "Mean squared error: Love it or leave it?," *IEEE Signal Processing Magazine*, pp. 98–117, 2009.

[82] Kalpana Seshadrinathan, Rajiv Soundararajan, Alan Conrad Bovik, and Lawrence K Cormack, "Study of subjective and objective quality assessment of video," *IEEE transactions on Image Processing*, vol. 19, no. 6, pp. 1427–1441, 2010.

[83] Stefan Winkler and Praveen Mohandas, "The evolution of video quality measurement: From psnr to hybrid metrics," *IEEE Transactions on Broadcasting*, vol. 54, no. 3, pp. 660–668, 2008.

[84] Kjell Brunnstrom, David Hands, Filippo Speranza, and Arthur Webster, "Vqeg validation and itu standardization of objective perceptual video quality metrics [standards in a nutshell]," *IEEE Signal processing magazine*, vol. 26, no. 3, 2009.

[85] Jiarun Song, Fuzheng Yang, Yicong Zhou, and Shan Gao, "Parametric planning model for video quality evaluation of iptv services combining channel and video characteristics," *IEEE Transactions on Multimedia*, 2016.

[86] Zhibo Chen, Ning Liao, Xiaodong Gu, Feng Wu, and Guangming Shi, "Hybrid distortion ranking tuned bitstream-layer video quality assessment," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 6, pp. 1029–1043, 2016.

[87] Shyamprasad Chikkerur, Vijay Sundaram, Martin Reisslein, and Lina J Karam, "Objective video quality assessment methods: A classification, review, and performance comparison," *IEEE transactions on broadcasting*, vol. 57, no. 2, pp. 165–182, 2011.

[88] Stefan Winkler, *Digital video quality: vision models and metrics*, John Wiley & Sons, 2005.

[89] Hong Ren Wu and Kamisetty Ramamohan Rao, *Digital video image quality and perceptual coding*, CRC press, 2005.

[90] "Msu software," http://www.compression.ru/video/quality_measure/video_measurement_tool.html.

[91] Zhou Wang, Ligang Lu, and Alan C Bovik, "Video quality assessment based on structural distortion measurement," *Signal processing: Image communication*, vol. 19, no. 2, pp. 121–132, 2004.

[92] Zhou Wang, Eero P Simoncelli, and Alan C Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. Signals, Systems and Computers*, 2003, vol. 2, pp. 1398–1402.

[93] Margaret H Pinson and Stephen Wolf, "A new standardized method for objectively measuring video quality," *IEEE Transactions on broadcasting*, vol. 50, no. 3, pp. 312–322, 2004.

[94] "Video quality metric software," https://www.its.bldrdoc.gov/resources/video-quality-research/software.aspx.

[95] Kalpana Seshadrinathan and Alan Conrad Bovik, "Motion tuned spatio-temporal quality assessment of natural videos," *IEEE transactions on image processing*, vol. 19, no. 2, pp. 335–350, 2010.

[96] "Motion-based video integrity evaluation (movie) index," http://live.ece.utexas.edu/research/quality/movie.html.

[97] Phong V Vu and Damon M Chandler, "Vis3: an algorithm for video quality assessment via analysis of spatial and spatiotemporal slices," *Journal of Electronic Imaging*, vol. 23, no. 1, pp. 013016–013016, 2014.

[98] "Vis3 source code," http://vision.eng.shizuoka.ac.jp/vis3/.

[99] Abdul Rehman, Kai Zeng, and Zhou Wang, "Display device-adapted video quality-of-experience assessment," in *Proc. SPIE/IS&T Electronic Imaging*, 2015.

[100] "Ssimplus software," http://www.ssimwave.com/.

[101] "Netflix techblog," .

[102] "Perceptual video quality assessment based on multi-method fusion," https://github.com/Netflix/vmaf.

[103] Kristian Skarseth, Henrik Bjørlo, Pål Halvorsen, Michael Riegler, and Carsten Griwodz, "Openvq: A video quality assessment toolkit," in *Proc. ACM*, 2016, pp. 1197–1200.

[104] "Openvq toolkit," https://bitbucket.org/mpg_code/openvq.

[105] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[106] Margaret H Pinson and Stephen Wolf, "Video quality model for variable frame delay (vqm_vfd)," Tech. Rep., NTIA Technical Memorandum TM-11-482, 2011.

[107] Eric C Larson and Damon M Chandler, "Most apparent distortion: full-reference image quality assessment and the role of strategy," *Journal of Electronic Imaging*, vol. 19, no. 1, 2010.

[108] Hamid R Sheikh and Alan C Bovik, "Image information and visual quality," *IEEE Transactions on Image Processing*, vol. 15, no. 2, pp. 430–444, 2006.

[109] Songnan Li, Fan Zhang, Lin Ma, and King Ngi Ngan, "Image quality assessment by separately evaluating detail losses and additive impairments," *IEEE Transactions on Multimedia*, vol. 13, no. 5, pp. 935–949, 2011.

[110] T Vlachos, "Detection of blocking artifacts in compressed video," *Electronics Letters*, vol. 36, no. 13, pp. 1106–1108, 2000.

[111] Stefan Winkler, Animesh Sharma, and David McNally, "Perceptual video quality and blockiness metrics for multimedia streaming applications," in *Proceedings of the international symposium on wireless personal multimedia communications*, 2001, pp. 547–552.

[112] Yuanyi Xue, Beril Erkin, and Yao Wang, "A novel no-reference video quality metric for evaluating temporal jerkiness due to frame freezing," *IEEE Transactions on Multimedia*, vol. 17, no. 1, pp. 134–139, 2015.

[113] R Venkatesh Babu, Ajit S Bopardikar, Andrew Perkis, and Odd Inge Hillestad, "No-reference metrics for video streaming applications," in *Proc. International Workshop on Packet Video*, 2004.

[114] Vladimir Zlokolica, Dragan Kukolj, Nemanja Lukic, and Miodrag Temerinac, "Evaluation on the selection of video quality metrics for overall visual perception," in *Proc. QoMEX*, 2010.

[115] Dubravko Ćulibrk, Dragan Kukolj, Petar Vasiljević, Maja Pokrić, and Vladimir Zlokolica, "Feature selection for neural-network based no-reference video quality assessment," in *Proc. ICANN*, 2009, pp. 633–642.

[116] Savvas Argyropoulos, Alexander Raake, Marie-Neige Garcia, and Peter List, "No-reference video quality assessment for SD and HD H.264/AVC sequences based on continuous estimates of packet loss visibility," in *Proc. QoMEX*, 2011.

[117] Savvas Argyropoulos, Alexander Raake, Marie-Neige Garcia, and Peter List, "No-reference bit stream model for video quality assessment of H.264/AVC video based on packet loss visibility," in *Proc. ICASSP*, 2011.

[118] P Gray, MP Hollier, and RE Massara, "Non-intrusive speech-quality assessment using vocal-tract models," *IEE Proceedings-Vision, Image and Signal Processing*, vol. 147, no. 6, pp. 493–501, 2000.

[119] TS 103 281 ETSI, "Speech and multimedia transmission quality (stq); speech quality in the presence of background noise: Objective test methods for super-wideband and fullband terminals," 2017.

[120] TS 103 106 ETSI, "Speech quality performance in the presence of background noise: background noise transmission for mobile terminals-objective test methods," 2014.

[121] ITU-T Recommendation P.564, "Conformance testing for voice over ip transmission quality assessment models," November 2007.

[122] A Rix and P Gray, "Niqa-non-intrusive speech quality assessment," *Contribution UIT-T COM*, 2001.

[123] ETR ETSI, "250 (1996)," *Speech Communication Quality from Mouth to Ear for 3, 1 kHz Handset Telephony across Networks*, 1996.

[124] ITU-T Rec. G.107, "The e-model, a computational model for use in transmission planning," 2003.

[125] Haytham Assem, David Malone, Jonathan Dunne, and Pat O'Sullivan, "Monitoring voip call quality using improved simplified e-model," in *Proc. ICNC*, 2013, pp. 927–931.

[126] ITU-T Rec. G.107.1, "Wideband e-model," June 2015.

[127] Ramon Sanchez-Iborra, Maria-Dolores Cano, and Joan Garcia-Haro, "Revisiting VoIP QoE assessment methods: are they suitable for VoLTE?," *Network Protocols and Algorithms*, vol. 8, no. 2, pp. 39–57, 2016.

[128] ITU-T Rec. BT.1788, "Methodology for the subjective assessment of video quality in multimedia applications," 2007.

[129] N Château, "Study of the influence of experimental context on the relationships between audio, video and audiovisual subjective qualities," .

[130] Marie-Neige Garcia, Robert Schleicher, and Alexander Raake, "Impairment-factor-based audiovisual quality model for iptv: influence of video resolution, degradation type, and content type," *Journal on Image and Video Processing EURASIP*, vol. 2011, no. 1, pp. 629284, 2011.

[131] Junyong You, Ulrich Reiter, Miska M Hannuksela, Moncef Gabbouj, and Andrew Perkis, "Perceptual-based quality assessment for audio–visual services: A survey," *Signal Processing: Image Communication*, vol. 25, no. 7, pp. 482–501, 2010.

[132] John G Beerends and Frank E De Caluwe, "The influence of video quality on perceived audio quality and vice versa," *Journal of the Audio Engineering Society*, vol. 47, no. 5, pp. 355–362, 1999.

[133] Francesca De Simone, Matteo Naccari, Marco Tagliasacchi, Frederic Dufaux, Stefano Tubaro, and Touradj Ebrahimi, "Subjective quality assessment of H.264/AVC video streaming with packet losses," *EURASIP Journal on Image and Video Processing*, vol. 2011, no. 1, pp. 190431, 2011.

[134] Stefan Winkler and Christof Faller, "Perceived audiovisual quality of low-bitrate multimedia content," *Multimedia, IEEE Transactions on*, vol. 8, no. 5, pp. 973–980, 2006.

[135] Stefan Winkler and Frédéric Dufaux, "Video quality evaluation for mobile streaming applications," in *Proc. Visual Communications and Image Processing*. International Society for Optics and Photonics, 2003, pp. 593–603.

[136] Ralf Steinmetz, "Human perception of jitter and media synchronization," *IEEE Journal on Selected Areas in Communications*, vol. 14, no. 1, pp. 61–72, 1996.

[137] T. Hayashi K. Yamagishi and A. Takahashi, "Planning model for audiovisual communication services," *NTT Technical Review*, vol. 7, no. 4, Apr 2009.

[138] N. Egi K. Yamagishi and T. Tominaga, "Monitoring the quality of iptv services," *NTT Technical Review*, vol. 11, no. 5, May 2013.

[139] ITU-T Rec. H.323, "Packet-based multimedia communications systems," December 2009.

[140] "Netdisturb software," https://www.ffmpeg.org/.

[141] ITU-T Rec. P.1401, "Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models," July 2012.

[142] Dominic W Massaro, Michael M Cohen, and Paula MT Smeele, "Perception of asynchronous and conflicting visual and auditory speech," *Journal of the Acoustical Society of America*, vol. 100, no. 3, pp. 1777–1786, 1996.

[143] "Ffmpeg," https://www.zti-communications.com/netdisturb/, 2016.

[144] ITU-T Rec. P.800, "Methods for subjective determination of transmission quality," August 1996.

[145] ITU-T Rec. P.910, "Subjective video quality assessment methods for multimedia applications," April 2008.

[146] ITU-T Rec. P.911, "Subjective audiovisual quality assessment methods for multimedia applications," December 1998.

[147] ITU-T Rec. BT-500, "Methodology for the subjective assessment of the quality of television pictures," 2012.

[148] JG Beerends and FE de Caluwe, "Relations between audio, video and audiovisual quality," *Contr COM*, pp. 12–19, 1997.

[149] Mike P Hollier, Andrew N Rimell, David S Hands, and Rupert M Voelcker, "Multi-modal perception," *BT Technology Journal*, vol. 17, no. 1, pp. 35–46, 1999.

[150] Benjamin Belmudez and Sebastian Möller, "Audiovisual quality integration for interactive communications," *EURASIP Journal on Audio, Speech, and Music Processing*, , no. 1, pp. 1–23, 2013.

[151] L Mued, B Lines, S Furnell, and P Reynolds, "The effects of audio and video correlation and lip synchronization," *Campus-Wide Information Systems*, vol. 20, no. 4, pp. 159–166, 2003.

[152] Benjamin Belmudez, Sebastian Moeller, Blazej Lewcio, Alexander Raake, and Amir Mehmood, "Audio and video channel impact on perceived audio-visual quality in different interactive contexts," in *Proc. MMSP*, 2009, pp. 1–5.

[153] Takanori Hayashi, Kazuhisa Yamagishi, Toshiko Tominaga, and Akira Takahashi, "Multimedia quality integration function for videophone services," in *Proc. GLOBECOM*, 2007, pp. 2735–2739.

[154] Franz Bräuer, Muhammad Sarwar Ehsan, and Gernot Kubin, "Subjective evaluation of conversational multimedia quality in ip networks," in *Proc. MMSP*, 2008, pp. 872–876.

[155] Anna Hart, "Mann-whitney test is not just a test of medians: differences in spread can be important," *British Medical Journal*, vol. 323, no. 7309, pp. 391, 2001.

[156] ITU-R Rec. BT-500.13, "Methodology for the subjective assessment of the quality of television pictures," 2012.

[157] Jari Korhonen, Ulrich Reiter, and Eugene Myakotnykh, "On the relative importance of audio and video in the presence of packet losses," in *Proc. QoMEX*, 2010, pp. 64–69.

[158] Ralf Steinmetz, "Human perception of jitter and media synchronization," *Selected Areas in Communications, IEEE Journal on*, vol. 14, no. 1, pp. 61–72, 1996.

[159] Anush Krishna Moorthy, Lark Kwon Choi, Gustavo De Veciana, and Alan Conrad Bovik, "Mobile video quality assessment database," *IEEE ICC Workshop on Realizing Advanced Video Optimized Wireless Networks*, 2012.

[160] "Epfl database," http://vqa.como.polimi.it/.

[161] Fadi Boulos, Wei Chen, Benoît Parrein, and Patrick Le Callet, "Region-of-interest intra prediction for H.264/AVC error resilience," in *Proc. ICIP*, 2009, pp. 3109–3112.

[162] Yohann Pitrey, Ulrich Engelke, Marcus Barkowsky, Romuald Pépion, and Patrick Le Callet, "Aligning subjective tests using a low cost common set," in *Proc. Euro ITV*, Lisbonne, Portugal, June 2011, p. irccyn contribution.

[163] Yohann Pitrey, Marcus Barkowsky, Patrick Le Callet, and Romuald Pépion, "Evaluation of MPEG4-SVC for QoE protection in the context of transmission errors," in *Proc. SPIE Optical Engineering*, San Diego, United States, Aug. 2010.

[164] Yohann Pitrey, Ulrich Engelke, Patrick Le Callet, Marcus Barkowsky, and Romuald Pépion, "Subjective quality of svc-coded videos with different error-patterns concealed using spatial scalability," in *Proc. EUVIP*, Paris, France, July 2011.

[165] Yohann Pitrey, Ulrich Engelke, Marcus Barkowsky, Romuald Pépion, and Patrick Le Callet, "Aligning subjective tests using a low cost common set," in *Proc. Euro ITV*, Lisbonne, Portugal, June 2011.

[166] Ines Saidi, Lu Zhang, Vincent Barriac, and Olivier Deforges, "Interactive vs. non-interactive subjective evaluation of ip network impairments on audiovisual quality in videoconferencing context," in *Proc. Qomex*, 2016, pp. 1–6.

[167] Tobias Hossfeld, Christian Keimel, Matthias Hirth, Bruno Gardlo, Julian Habigt, Klaus Diepold, and Phuoc Tran-Gia, "Best practices for qoe crowdtesting: Qoe assessment with crowdsourcing," *IEEE Transactions on Multimedia*, vol. 16, no. 2, pp. 541–558, 2014.

[168] Tobias Hoßfeld, Matthias Hirth, Pavel Korshunov, Philippe Hanhart, Bruno Gardlo, Christian Keimel, and Christian Timmerer, "Survey of web-based crowdsourcing frameworks for subjective quality assessment," in *Proc. MMSP*, 2014, pp. 1–6.

[169] Flávio Ribeiro, Dinei Florêncio, Cha Zhang, and Michael Seltzer, "Crowdmos: An approach for crowdsourcing mean opinion score studies," in *Proc. ICASSP*, 2011, pp. 2416–2419.

[170] Christian Keimel, Julian Habigt, Clemens Horch, and Klaus Diepold, "Qualitycrowd—a framework for crowd-based quality evaluation," in *Proc. PCS*, 2012, pp. 245–248.

[171] Louis Anegekuh, Lingfen Sun, and Emmanuel Ifeachor, "A screening methodology for crowdsourcing video qoe evaluation," in *Proc. GLOBECOM*, 2014, pp. 1152–1157.

[172] Tobias Hossfeld, Matthias Hirth, Pavel Korshunov, Philippe Hanhart, Bruno Gardlo, Christian Keimel, and Christian Timmerer, "Survey of web-based crowdsourcing frameworks for subjective quality assessment," in *Proc. MMSP*, 2014, pp. 1–6.

[173] "The foulefactory web site," https://www.foulefactory.com.

[174] "ITU Contribution 0382: Assessing audio quality with a "crowdsourcing" approach and comparison with laboratory : second results and first recommendations, author= Laetitia crowd ," .

[175] VQEG, "Final report from the video quality experts group on the validation of objective models of multimedia quality assessment, phase i," 2008.

[176] Stephen Wolf and Margaret Pinson, "Video quality measurement techniques," *2002.*, 2002.

[177] Demóstenes Zegarra Rodríguez, Renata Lopes Rosa, and Graça Bressan, "No-reference video quality metric for streaming service using dash standard," in *Proc. ICCE*, 2015, pp. 106–107.

[178] M Slanina, V Ricny, and R Forchheimer, "A novel metric for H.264/AVC no-reference quality assessment," in *Proc. EURASIP*, 2007, pp. 114–117.

[179] Christian Keimel, Manuel Klimpke, Julian Habigt, and Klaus Diepold, "No-reference video quality metric for HDTV based on H.264/AVC bitstream features," in *Proc. ICIP*, 2011, pp. 3325–3328.

[180] Federica Battisti, Marco Carli, Yiwei Liu, Alessandro Neri, and Pradip Paudyal, "Distortion-based no-reference quality metric for video transmission over ip," in *Proc. ISSCS*, 2015, pp. 1–4.

[181] Federica Battisti, Marco Carli, and Alessandro Neri, "No reference quality assessment for mpeg video delivery over ip," *EURASIP Journal on Image and Video Processing*, , no. 1, pp. 13, 2014.

[182] Mikołaj Leszczuk, Mateusz Hanusiak, Mylène CQ Farias, Emmanuel Wyckens, and George Heston, "Recent developments in visual quality monitoring by key performance indicators," *Multimedia Tools and Applications*, vol. 75, no. 17, pp. 10745–10767, 2016.

[183] Piotr Romaniak, Lucjan Janowski, Mikolaj Leszczuk, and Zdzislaw Papir, "Perceptual quality assessment for H.264/AVC compression," in *Proc. CCNC*, 2012.

[184] Anish Mittal, Michele A Saad, and Alan C Bovik, "A completely blind video integrity oracle," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 289–300, 2016.

[185] ITU-R Rec. P.862, "Perceptual evaluation of speech quality (pesq): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," 2001.

[186] ITU-T Temporary Document TD 724 Rev.1 (GEN/12), "Technical requirement specification p.spelq," 2015.

[187] Muhammad Sajid Mushtaq and Abdelhamid Mellouk, *Quality of Experience Paradigm in Multimedia Services: Application to Ott Video Streaming and Voip Services*, Elsevier, 2017.

[188] "Khiops data mining tool web site," https://khiops.predicsis.com/.

[189] Pedro Domingos and Michael Pazzani, "On the optimality of the simple Bayesian classifier under zero-one loss," *Machine learning*, vol. 29, no. 2, pp. 103–130, 1997.

Résumé de la thèse en Français

1. Thématique et enjeux

Avec les progrès technologiques, les services de télécommunication évoluent en intégrant de nouvelles modalités au-delà du simple échange vocal bien connu de tous. Ainsi voit-on de plus en plus de services mêlant le son, le texte et la vidéo. C'est le cas des applications de messagerie instantanée, où les fonctions classiques de messagerie sont combinées avec de la voix sur IP, voire même de la visiophonie (Skype, Messenger, Google Duo, iChat…). Quel que soit le contexte d'utilisation, personnel ou professionnel, on est passé du simple appel téléphonique ou la réactivité visuelle est absente aux appels vidéo. Selon les dernières statistiques reportées par CISCO [1], les services vidéo mobiles vont générer trois quarts du trafic des données mobiles en 2020. L'image devient un élément indispensable pour bien communiquer avec les autres. Cependant, plusieurs facteurs influencent la qualité de nos appels vidéo et donc impactent notre qualité d'expérience. Les méthodes de mesure de la qualité perçue des services conversationnels audiovisuels ne sont pas encore matures et exploitées par les opérateurs de télécommunication pour plusieurs raisons :

- usage presque exclusif à l'internet,
- complexité technique de la mesure de qualité vidéo
- multiplicité et complexité des terminaux et écrans utilisés (PC, smartphone, TV, etc.)

C'est pourquoi l'identification de méthodes adéquates pour la mesure et la supervision de la qualité perçue de ces nouveaux services devient un défi majeur pour les opérateurs de télécommunications. Dans ce contexte, l'enjeu de la thèse est d'étudier et de proposer des métriques représentatives de la perception de la qualité des flux associés aux services de visiophonie et visioconférence. Ces métriques seront à déterminer à partir d'informations issues du signal audio et vidéo, mais aussi d'éléments d'analyse du fonctionnement du service, accessibles au niveau du terminal ou d'équipements de réseau. Des tests subjectifs sont menés afin de collecter le jugement des utilisateurs de services sur la qualité perçue en fonction de différents niveaux de dégradations. Le principe général est ensuite d'établir une corrélation forte entre les métriques objectives sélectionnées et la qualité perçue telle qu'elle est exprimée par les utilisateurs.

Les résultats de mes travaux de recherche servent à mettre en place une boite d'outils composée des métriques représentatives de qualité vidéo. Cette boite permet de faire le monitoring et le diagnostic de la qualité d'un service audiovisuel en identifiant à partir d'un

ensemble de métriques objectives le type des dégradations présentes (perte de bloc, pixellisation, flou, gel d'image, …etc.) et leurs possibles causes.

2. Objectifs de la thèse

L'objectif de la thèse est d'étudier et de proposer des métriques représentatives de la perception de la qualité des flux associés aux services de visiophonie et visioconférence. Ces métriques seront à déterminer à partir d'informations issues du signal audio et vidéo.

Nous allons nous appuyer sur les domaines de la mesure et de la modélisation de qualité vidéo, audio et audiovisuelle perçue existants. Nous sommes intéressés par les métriques liées à l'analyse du signal (notes MOS, estimation de la qualité globale, détection d'artefacts perçus).

Parmi les objectifs c'est de fournir à l'opérateur la connaissance, en termes d'impact sur la qualité perçue, des mécanismes présents dans les réseaux et terminaux supportant les services conversationnels audiovisuels, susceptibles de perturber la qualité perçue de ces services. Au-delà des besoins de l'opérateur, cette connaissance a vocation à être partagée avec l'écosystème, pour aboutir au développement de nouveaux outils de supervision.

3. Etat de l'Art

La première partie de la thèse a été consacrée à un état de l'art exhaustif sur :
- les méthodes de tests subjectifs applicables au contexte d'évaluation de la qualité audiovisuelle. Notamment les recommandations ITU P.800 [2], ITU P.910 [3] et ITU P.911 [4],
- les méthodes et modèles objectifs existants d'évaluation de la qualité vidéo, audio et audiovisuelle.

3.1. Méthodes subjectives d'évaluation de la qualité audiovisuelle

Les évaluations subjectives constituent le moyen le plus précis de mesurer la qualité d'un flux multimédia. Dans les tests subjectifs, un certain nombre de sujets (observateurs ou participants) sont invités à assister à un ensemble de tests et à juger de la qualité des médias ou de l'inconvénient causé par les distorsions. La moyenne des valeurs obtenues pour chaque séquence de test est connue sous le nom de Mean Opinion Score (MOS). En général, les évaluations subjectives sont coûteuses et prennent beaucoup de temps. En conséquence, le nombre test pouvant être réalisés est limité et, par conséquent, une méthodologie appropriée doit être utilisée pour utiliser au mieux les ressources. Dans le contexte des services de vidéoconférence, dans les systèmes multimédias généraux, la Recommandation UIT-T P.910 fournit des méthodes d'évaluation de la qualité vidéo. Seules deux normes sont consacrées à

l'évaluation subjective de la qualité audiovisuelle pour un contexte interactif (UIT-T P.920 [5]) ou non-interactif (UIT-T P.911).

3.2. Modèle objectif d'évaluation de la qualité audiovisuelle

La recommandation UIT-T G.1070 [6] décrit un modèle de calcul paramétrique pour les applications de visiophonie point à point sur les réseaux IP normalisés par l'UIT en 2012. L'algorithme estime la qualité perçue sur la base de paramètres de mesure, mais non sur la base des signaux vidéo et audio. Les entrées du modèle sont des informations sur le codec, le débit codé, les erreurs de transport et les informations sur la mise en mémoire tampon.

L'algorithme est conçu pour estimer la qualité d'un contenu audiovisuel typique et moyen et donner le même score pour un codec donné, un débit binaire et une situation d'erreur de transport indépendants du contenu audiovisuel.

Cet algorithme paramétrique est capable d'évaluer la qualité du flux audiovisuel en direct, puisque des informations détaillées sur la vidéo source ne sont pas nécessaires. L'algorithme nécessite généralement des informations sur le codec et le débit binaire codé. Ce type d'algorithme peut toujours être applicable lorsque seul un flux binaire crypté est disponible.

Le modèle G.1070 est composé de trois modules de qualité : les modules audio, vidéo et audiovisuel. En sortie, les modules fournissent une estimation individuelle des qualités audio et vidéo et le modèle les combine tous dans une fonction d'intégration pour la qualité audiovisuelle globale sur l'échelle ACR à 5 points.

3.3. Métriques objectives d'évaluation de la qualité vidéo

3.3.1. Métriques avec référence complète

Dans l'évaluation de la qualité vidéo, les métriques avec référence complète effectuent une comparaison entre un flux vidéo de référence et un flux vidéo dégradé. Dans ce type d'approche, nous supposons que la perte de qualité est directement liée à un signal d'erreur ajouté à un signal initialement "Parfait". Étant donné que ce type de métrique nécessite la disponibilité de la totalité de la vidéo de référence, ils ne sont pas utiles pour l'évaluation en temps réel et la surveillance. Les métriques avec référence imposent généralement un alignement spatial et temporel précis des deux signaux. Dans le tableau 1 nous présentons un récapitulatif des métriques que nous avons étudiées.

3.3.2. Métriques sans référence

Métriques de MOAVI

Ces métriques de la qualité audiovisuelle sont développées par le Département des Télécommunications à l'Université des Sciences et Technologies AGH. Ce travail de recherche fait partie du projet MOAVI (Monitoring de la Qualité Audiovisuelle par les Indicateurs Clés) au sein du Groupe d'Experts en Qualité Vidéo (VQEG). Les mesures proposées évaluent la présence de différentes dégradations de la qualité vidéo telles que la pixellisation, la perte de bloc, le flou, le bruit, le scintillement, etc.

Métrique	Approche	Intervalle de valeurs	Temps d'exécution	Implémentation
PSNR	Mesure de l'erreur quadratique moyenne	[0,100]	1	Logiciel MSU
SSIM [7]	Mesure de la distorsion structurée	[0,1]	1,05	Logiciel MSU
MS-SSIM [8]	Mesure de la distorsion structurée multi échelle	[0,1]	2	Logiciel MSU
VQM [9]	Filtre de dégradation des contours	[0,1]	30	Logiciel NTIA
MOVIE [10]	Filtre Gabor	[0,1]	456	Code source
ViS3 [11]	Algorithme MAD	[0,100]	23	Matlab code source
SSIMplus [12]	Fonction de sensitivité au contraste	[0,100]	4	Logiciel SSIMwave
VMAF [13]	Apprentissage par Machine Learning	[0,100]	26	Code source
OPVQ [14]	ITU-T J.247	[1,5]	19	OpenVQ Toolkit

Nous avons sélectionné un ensemble de mesures sans référence que nous jugeons représentatives des types de dégradation susceptibles d'infecter une vidéoconférence ou un appel de téléphonie vidéo. Par conséquent, dans notre étude, nous avons mis l'accent sur l'étude de l'effet des artefacts de traitement et de transmission sur la qualité perçue par les utilisateurs de services vidéo.

La métrique VIIDEO

C'est une métrique de qualité vidéo complètement aveugle qui n'exige pas la présence de la vidéo de référence ou des jugements humains pour la formation [15]. La métrique ne modélise aucune information spécifique à la distorsion, mais ne fait que modéliser le «caractère naturel» statistique (ou son absence) de la vidéo. L'algorithme est basé sur les corrélations entre sous-bandes pour quantifier le degré de distorsion présent dans la vidéo et ainsi prédire les jugements humains de qualité vidéo. De plus, la complexité temporelle de chaque étape de l'algorithme d'évaluation de l'intégrité intrinsèque et de la distorsion est analysée. La métrique de VIIDEO suppose que pour une vidéo de bonne qualité, ses statistiques locales de différences de trames traitées par la suppression locale moyenne et la normalisation de contraste de division devraient suivre une distribution gaussienne généralisée.

4. Tests subjectifs : analyses expérimentales

Nos études subjectives ont deux objectifs : évaluer la perception des utilisateurs de services de vidéoconférence dans différentes conditions et constituer une base de données de séquences pour évaluer la performance des métriques de qualité objective. Nous étudions la qualité vidéo, audio et audiovisuelle et la perception de la désynchronisation dans deux situations différentes : une conversation interactive et un test non interactif. Nous analysons les effets des dégradations de réseau (perte de paquets, retard) sur la qualité audiovisuelle perçue, audio et vidéo. Nous évaluons l'impact du contexte expérimental et de la complexité des scènes sur la perception de la qualité en cas d'appels vidéo. De plus, nous proposons de nouveaux seuils d'acceptabilité de la désynchronisation audio-vidéo dans le contexte de la visiophonie et étudions l'effet de la synchronisation en présence et en l'absence de dégradation du réseau.

Nous avons étudié deux modalités de test de qualité audiovisuelle subjective : dans des contextes interactifs et non interactifs et sous différentes complexités de scène. En comparant les résultats de tests non interactifs et interactifs, nous résumons statistiquement qu'il n'y a pas de différence significative dans la perception de la qualité audio, vidéo et la perception de la désynchronisation entre les deux contextes expérimentaux. Ainsi, dans les expériences futures, nous pouvons nous appuyer sur des résultats de test non interactifs et les appliquer

dans un contexte conversationnel. Cependant, en considérant la perception de la qualité audiovisuelle, nous notons une différence significative entre les deux contextes.

Par ailleurs, les résultats montrent que la complexité de la scène a un impact sur la qualité audiovisuelle perçue dans les deux contextes et sur la perception de la synchronisation audio-vidéo dans le contexte interactif. L'observation différente sur l'impact de la complexité de la scène sur la qualité de la vidéo dans les deux contextes nécessite une étude plus approfondie. Limitée par la durée de l'expérience nous avons étudié seulement deux scènes différentes dans le contexte interactif. Nous n'avions pas couvert un large éventail de complexité spatiale et temporelle.

La réduction de l'alignement temporel entre l'information auditive et visuelle peut altérer la perception audiovisuelle en tant qu'événement multimodal. Dans la conversation audiovisuelle en temps réel, la présence d'une désynchronisation entre l'image et le son peut avoir un effet néfaste sur l'interactivité de la conversation et donc sur la qualité perçue. Par conséquent, il est nécessaire de contrôler la relation temporelle entre les signaux audio et vidéo afin que la qualité perçue par l'utilisateur ne soit pas altérée.

Nous avons étudié l'impact de la désynchronisation sur la perception de la qualité audiovisuelle dans le contexte du contenu de la vidéoconférence. Les résultats ont montré qu'il existe le même aspect de dissymétrie pour les applications TV et visiophonie, mais avec des seuils d'acceptabilité plus grands, au moins 150 et 250 ms respectivement. La raison de cette différence n'est pas nécessairement liée au contexte ; la conception des tests subjectifs respectifs, et en particulier la question posée, peut également avoir un impact. C'est pourquoi nous avons réalisé un nouveau test subjectif incluant des valeurs de désynchronisation plus élevées, afin de voir si les seuils d'acceptabilité réels pourraient être encore plus élevés.

De plus, nous nous intéressons à l'interaction entre différents types de dégradations et à la synchronisation audio / vidéo pouvant conduire à des effets de masquage visuel. En particulier, nous voulons étudier si le passage d'une haute résolution à une faible résolution ou si le débit binaire du codec vidéo peut avoir un impact sur la perception de l'asynchronisme par l'utilisateur. L'interaction entre la désynchronisation et la perte de paquets (audio ou vidéo) peut également conduire à un masquage visuel. Ainsi, nous donnerons des éléments de réponse à cette problématique qu'elle n'est pas encore étudiée dans la littérature.

Les résultats de ce test subjectif ont montré que la complexité spatiale des contenus vidéo et le bruit des contextes audio ont une influence négative sur la perception de la désynchronisation (plus complexe ou plus bruyant, pire en termes de perception), En plus, en présence de perte d'audio d'information vidéo (résultant de la perte de paquets IP), la désynchronisation est moins perceptible.

Ces éléments de connaissance devraient être pris en compte dans le développement de toutes les futures méthodes subjectives et objectives de QoE qui traitent de l'évaluation de la qualité perçue des services conversationnels audiovisuels.

5. Evaluation des métriques objectives

La littérature propose de nombreuses méthodes d'évaluation objective des qualités audio, vidéo et audiovisuelles. L'objectif de cette partie est d'évaluer la performance des principales approches et mesures existantes pour prédire la qualité vidéo, et audiovisuelle. Les applications de l'évaluation objective de la qualité sont diverses. Le post-traitement, la transmission, les capteurs ou les affichages sont des éléments pouvant être soumis à des critères de qualité spécifiques. Notre contribution principale est d'étudier la performance des modèles objectifs en fonction des différentes dégradations qui peuvent survenir par exemple lors d'une conférence vidéo.

5.1. Métriques de qualité vidéo avec référence complète

Nous avons effectué une comparaison des performances de dix métriques objectives (voir Tableau 1) différentes dans le contexte de l'appel vidéo et de la vidéoconférence. La comparaison des mesures a été effectuée en fonction de leur précision de prédiction, de leur monotonie et de leur stabilité. Dans cette étude, nous avons utilisé deux bases de données de qualité vidéo publique (EPFL et LIVE Mobile) et deux bases de données créées dans le cadre de nos tests de qualité subjective de visioconférence audiovisuelle chez Orange Labs. Les résultats expérimentaux montrent que les métriques qui incluent des informations sur l'aspect temporel de la vidéo dans l'algorithme d'estimation de la qualité dépassent les autres métriques. Pour la base de données EPFL qui ne contient que les erreurs de transmission de perte de paquets, toutes les métriques sont bien corrélées avec la perception subjective de la qualité vidéo, avec une petite préférence pour OPVQ, Vis3, SSIMplus et VMAF. Pour le même type de dégradation avec des contenus plus proches de ceux dans le contexte de la vidéoconférence, ViS3 surpasse statistiquement les autres métriques testées.

En ce qui concerne les dégradations causées par les débits de codage H.264 et HEVC, VMAF et SSIMplus sont les métriques les plus compétitives. Pour une base de données de types de dégradation croisés, OPVQ, VMAF, ViS3 et SSIMplus ont des performances statistiques équivalentes à celles des autres métriques. Ainsi, les résultats expérimentaux montrent qu'il n'y a pas de métrique universelle qui soit la meilleure pour tous les types et tous les contenus de distorsion. Pour évaluer l'influence du type de codec, le débit de codage et les changements de fréquence d'images, OPVQ, ViS3, SSIMplus et VMAF peuvent donner des scores objectifs mieux corrélés avec le MOS. Cependant, d'autres études sont nécessaires pour optimiser l'algorithme OPVQ pour la nouvelle génération de codecs vidéo tels que le HEVC. Dans le cas d'erreurs de transmission réseau, nous avons une forte probabilité d'obtenir un

désalignement temporel entre la séquence de référence et la séquence dégradée. En conséquence, les scores de métriques basés sur la comparaison image par image sont biaisés. Dans ce cas, nous recommandons l'utilisation de la métrique ViS3 car son algorithme est basé sur la qualité informatique du GOP et du STS. VMAF est un modèle prometteur pour la qualité vidéo car il est construit en utilisant l'approche d'apprentissage automatique. Ses performances peuvent être améliorées en enrichissant l'ensemble de données d'apprentissage avec de grands types simples de déficiences et de contenus, et en formant d'autres métriques plus objectives telles que SSIMplus, ViS3 ... etc.

5.2. Métriques de qualité vidéo sans référence

Nous avons présenté une étude d'évaluation des performances de six métriques d'évaluation de la qualité vidéo développées par le projet MOAVI VQEG. L'étude a impliqué trois bases de données de test avec un large échantillon de types de dégradations. Nous trouvons que les mesures peuvent être des indicateurs représentatifs de la qualité vidéo. Pour chaque condition (codage, perte de paquets, affaiblissement de signal, etc.), nous avons identifié les métriques représentatives que nous recommandons de prendre en compte. Selon les résultats obtenus, on peut voir que pour les dégradations de transmission, les distorsions perçues par l'utilisateur final peuvent se manifester par des événements de perte de bloc, de découpage ou de congélation. En ce qui concerne les dégradations liées à l'encodage, elles sont essentiellement du flou, de la pixellisation et du scintillement. Ces métriques constituent une partie de la boîte à outils pour diagnostiquer la qualité vidéo dans les services de communication.

5.3. Evaluation du modèle G.1070

Pendant des années, l'Union Internationale des Télécommunications est intéressée par l'étude des aspects de Qualité de Service (QoS) et de Qualité de l'Expérience (QoE) pour le streaming multimédia et les services de communication. Le groupe d'étude chargé des Recommandations pour la QoE à l'UIT est le SG12 (QoS et QoE). En particulier, ce groupe d'étude travaille sur un modèle de planification de la qualité (G.1071) et sur des modèles de surveillance de la qualité vidéo et audiovisuelle (série P.120x) des applications de diffusion en continu. En ce qui concerne les applications de vidéo téléphonie, la seule norme existante est la Recommandation UIT-T G.1070 "Modèle d'opinion pour les applications de vidéo-téléphonie" (2012). Dans cette section, nous étudions la précision de la prédiction et la pertinence de ce modèle, initialement destinées à des fins de planification seulement.

Nous avons évalué le modèle G.1070 en nous basant sur nos résultats subjectifs interactifs et non interactifs.

Les résultats de corrélation entre les scores calculés par le modèle et les notes MOS d'évaluation subjective ont montré que les modules audiovisuels et audio ont des

performances bien inférieures à celles du module vidéo. Ce résultat peut s'expliquer par le fait que les modules audio et audiovisuels prennent comme paramètres d'entrée les retards de la parole et de la vidéo, contrairement au module vidéo. Nous remarquons que toutes les conditions où l'erreur entre la sortie du modèle G.1070 et le score subjectif est importante, sont les conditions avec un retard de la parole. Ainsi, on peut signaler que le modèle G.1070 sous-estime la qualité audio et audiovisuelle en cas de retard audio et considère que cette dégradation détériore la qualité plus largement que celle perçue par les sujets. Si nous ignorons les conditions de délai audio et que nous calculons la corrélation entre la métrique du modèle et les scores subjectifs, nous trouvons des résultats de corrélation meilleurs.

Si l'on compare avec les résultats de corrélation avec toutes les conditions, il est clair que ce modèle fournit une bonne estimation de la qualité subjective concernant la perte de paquets et le retard vidéo. Pour les bases de données subjectives non interactives et interactives, nous obtenons les mêmes résultats. Cela peut indiquer que ce type de scénario de test n'a pas d'effet sur le processus d'estimation de la qualité.

Sur ces études d'évaluation nous avons proposé des solutions d'amélioration du modèle sous la forme d'une contribution à l'Union Internationale de Télécommunication.

6. Application de l'approche de Machine Learning pour la génération d'un modèle global de qualité vidéo

L'évaluation de la qualité vidéo est une tâche complexe étant donné la multiplicité des paramètres ayant une incidence sur les médias perçus. La méthodologie des tests subjectifs d'évaluation de la qualité, bien qu'elle donne la perception exacte de la qualité, n'a pas pu être utilisée en temps réel. D'autre part, nous avons montré dans la section précédente que les outils et les modèles objectifs sont nombreux et qu'il n'y a pas de métrique représentative pour toutes les conditions de dégradation.

Dans notre contexte d'étude des services de visioconférence et de visiophonie, nous avons montré à travers nos tests subjectifs que la qualité audiovisuelle globale est généralement plus influencée par la qualité vidéo que par la qualité audio. C'est pourquoi nous nous concentrons principalement sur l'évaluation de la qualité vidéo d'un service de vidéoconférence en temps réel. Dans ce cas, nous considérons les métriques sans référence car en temps réel, le signal de référence de l'application n'est pas disponible. Chacune de ces mesures permet de mesurer le niveau d'un seul type de distorsion affectant un signal vidéo. Cependant, la perception humaine de la qualité ne fait pas de distinction entre les types de distorsion mais donne une appréciation globale de la qualité. Notre idée est alors d'essayer de combiner toutes les métriques basées sur des artefacts uniques de MOAVI dans un modèle de qualité vidéo global généré par des méthodes de Machine Learning (ML).

Machine Learning (ML) consiste en la conception et le développement de programmes et d'algorithmes qui ont la capacité d'améliorer automatiquement leur performance sur la base de leur propre expérience au fil du temps, ou de données antérieures fournies par d'autres programmes. Les fonctions générales fournies par ML sont l'entraînement, la reconnaissance, la généralisation, l'adaptation, l'amélioration et l'intelligibilité. Il existe deux types de ML, c'est-à-dire l'apprentissage non supervisé et supervisé. L'algorithme ML non supervisé trouve la structure cachée dans les données non étiquetées afin de les classer en catégories significatives, tandis que l'apprentissage supervisé suppose que la structure de catégorie ou la hiérarchie de la base de données est déjà connue. La ML supervisée nécessite un ensemble de classes étiquetées et renvoie une fonction qui mappe la base de données sur les étiquettes de classes prédéfinies. Il fait des prédictions sur les instances futures afin de construire un modèle concis qui représente la distribution des données. Dans notre cas, nous considérons l'apprentissage supervisé, et nous sommes intéressés par les méthodes de classification en raison de la nature discrète et étiquetée de notre ensemble de données et parce que notre objectif est de prédire une variable.

Nous avons étudié la possibilité de combiner des mesures d'artefacts uniques sans référence provenant de MOAVI dans un modèle global d'évaluation de la qualité vidéo. Le modèle obtenu a une précision de seulement 0,44 ce qui n'est pas suffisant pour un bon modèle. Après l'ajout d'aucune métrique de référence VIIDEO aux variables d'apprentissage de l'algorithme ML, le modèle est amélioré et atteint 0,63 de précision. Ce résultat est encourageant car nous considérons que même si notre base de données ne contient que 1130 séquences, ce volume a permis de générer un modèle de prédiction prometteur. Nous recommandons de collecter plus de bases de données avec des conditions plus diversifiées.

7. Conclusion et perspectives

Le travail réalisé dans cette thèse a conduit à plusieurs résultats dans le domaine de la QoE dans le cadre des services de visiophonie et de visioconférence. Les contributions sont doubles, car elles se rapportent à l'évaluation subjective et objective de la qualité audiovisuelle d'un appel vidéo. La première contribution est la constitution d'une base de données de séquences audiovisuelles correspondant à un scénario réel d'appel vidéo, une question cruciale pour la communauté de la qualité audiovisuelle. La deuxième contribution concerne l'évaluation des outils d'évaluation de la qualité objective existants.

En général, notre travail a permis de mieux comprendre les processus d'évaluation de la qualité audiovisuelle pour les services de visiophonie. Néanmoins, il reste encore quelques zones grises à éclaircir et la possibilité d'approfondir certaines des approches proposées. Plus important encore, l'élargissement de la base des séquences audiovisuelles permettrait un

meilleur apprentissage des critères objectifs. Cela permettrait également de réduire les inexactitudes sur les indicateurs de performance.

Dans tous nos tests subjectifs, nous nous sommes limités à l'évaluation du type d'application et des dégradations de transmission. Il est évident qu'un service de visiophonie est influencé par d'autres facteurs, tels que le contexte, la situation psychologique, le type de terminal, l'OS ... L'élargissement à un spectre plus large de déficiences et conditions permettrait une caractérisation plus fine de la qualité d'une vidéo Service téléphonique.

Plusieurs travaux supplémentaires sont réalisables sur les critères de qualité objectifs, en particulier dans le développement de solutions temps réel. Nous croyons que l'approche Machine Learning est prometteuse. Il est possible de collecter une base de données de formation de qualité vidéo plus importante afin de couvrir toutes les dégradations possibles de la qualité vidéo. De plus, nous n'avons formé notre modèle que sur les métriques MOAVI sans référence et sur la métrique VIIDEO. Il serait intéressant d'étudier d'autres variables telles que la résolution vidéo, le débit binaire de codage, le pourcentage de perte de paquets, etc. Ces variables apportent des informations supplémentaires à l'algorithme et le rendent plus décisif.

Liste des publications

I. Saidi, Lu Zhang, V. Barriac and O. Deforges, "Interactive vs. non-interactive subjective evaluation of IP network impairments on audiovisual quality in videoconferencing context," *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*, Lisbon, 2016, pp. 1-6.

I. Saidi, L. Zhang, O. Deforges, V. Barriac. “Evaluation of the performance of ITU-T G.1070 model for packet loss and desynchronization impairments”. QoMEX, June 2016, Lisbon, Portugal.

I. Saidi, L. Zhang, V. Barriac and O. Deforges, "Audiovisual quality study for videoconferencing on IP networks," *2016 IEEE 18th International Workshop on Multimedia Signal Processing (MMSP)*, Montreal, QC, 2016, pp. 1-6.

I. Saidi, L. Zhang, V. Barriac and O. Deforges, "Evaluation of single-artifact based video quality metrics in video communication context," *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*, Erfurt, 2017, pp. 1-3.

UIT-T SG 12, Evaluation of the performance of ITU-T G.1070 model for packet loss and desynchronization impairments, 2016

UIT-T SG 12, Results and analysis of two passive subjective tests of audiovisual quality of videotelephony contents in presence of asynchronism between sound and image conducted following respectively the P.911 procedure and a crowdsourcing approach, 2017

Bibliographie

[1] Cisco Visual Networking Index Cisco, “Global mobile data traffic forecast update, 2015–2020 white paper, 2016”

[2] ITU-T Recommendation P.800, _Methods for subjective determination of transmission quality, August 1996.

[3] ITU-T Recommendation P.910, _Subjective video quality assessment methods for multimedia applications, April 2008.

[4] ITU-T Recommendation P.911, _Subjective audiovisual quality assessment methods for multimedia applications, December 1998

[5]Recommendation ITU-TP.920, Interactive test methods for audiovisual communications, (ITU-T), May2000

[6] ITU-T Rec. G.1070, _Opinion model for video-telephony applications, Aug. 2012.

[7] Zhou Wang, Ligang Lu, and Alan C Bovik, “Video quality assessment based on structural distortion measurement,” Signal processing: Image communication, vol.19,no.2,pp.121132,2004.

[8] Zhou Wang, Eero P Simoncelli, and Alan C Bovik, “Multi scale structural similarity for image quality assessment,” in Signals, Systems and Computers, 2004.

[9] Margaret H Pinson and Stephen Wolf, “A new standardized method for objectively measuring video quality,” IEEE Transactions on broadcasting, vol. 50, no.3,pp.312_322,2004

[10] Kalpana Seshadrinathan and Alan Conrad Bovik, “Motion tuned spatiotemporal quality assessment of natural videos,” IEEE transactions on image processing, vol.19,no.2,pp.335_350,2010

[11] Phong V Vu and Damon M Chandler,”Vis3:an algorithm for video quality assessment via analysis of spatial and spatiotemporal slices,” Journal of Electronic Imaging, vol.23,no.1,pp.013016_013016,2014.

[12] Abdul Rehman, Kai Zeng, and Zhou Wang, “Display device-adapted video quality-of-experience assessment,” in SPIE/IS&T Electronic Imaging. International Society for Optics and Photonics, 2015,pp.939406_939406

[13] Netflix techblog, http://techblog.net ix.com/2016/06/toward-practical- perceptual-video.html

[14] Kristian Skarseth, Henrik Bjørlo,Pål Halvorsen, Michael Riegler,and Carsten Griwodz, “Openvq: A video quality assessment toolkit”, in Proceedings of the 2016 ACM on Multimedia Conference. ACM,2016,pp.1197_1200.

[15] Anish Mittal, Michele A Saad, and Alan C Bovik, “A completely blind video integrity oracle”, IEEE Transactions on Image Processing, vol.25,no.1,pp. 289_300, 2016.

INSA de RENNES
Service des Formations

AVIS DU JURY SUR LA REPRODUCTION DE LA THESE SOUTENUE

Titre de la thèse:
Analyse et modélisation de la qualité perçue des applications de visiophonie

Nom Prénom de l'auteur : SAIDI INES

Membres du jury :
- Monsieur DEFORGES Olivier
- Madame ZHANG Lu
- Monsieur LARABI Chaker
- Monsieur PUECH William
- Monsieur DUFAUX Frédéric
- Monsieur RAAKE Alexander
- Monsieur BARRIAC Vincent

Président du jury : W. PUECH

Date de la soutenance : 28 Février 2018

Reproduction de la these soutenue

☒ Thèse pouvant être reproduite en l'état
☐ Thèse pouvant être reproduite après corrections suggérées

Fait à Rennes, le 28 Février 2018

Signature du président de jury

William PUECH

Le Directeur,

M'hamed DRISSI

Résumé

Dans un contexte fortement concurrentiel, l'un des principaux enjeux pour les opérateurs et les fournisseurs de services de visiophonie est de garantir aux utilisateurs une qualité d'expérience (QoE) optimale. Il existe un fort besoin d'une mesure qui reflète la satisfaction et la perception des utilisateurs de ces services. La qualité audiovisuelle d'un appel vidéo doit être contrôlée pour répondre à deux besoins principaux. Le premier concerne la planification de nouvelles technologies en cours de développement. Le second est axé sur le contrôle des communications existantes en évaluant la qualité des services offerts.

Aujourd'hui, deux approches sont utilisées pour évaluer la qualité audiovisuelle : les tests subjectifs en collectant des notes données par des participants sur des échelles de qualité, après visualisation et écoute de séquences audiovisuelles et les métriques objectives basées sur des algorithmes automatiques d'évaluation de la qualité d'un signal audio, vidéo ou audiovisuel. Concernant les services de téléphonie, des décennies de recherche, de standardisation et d'exploitation des réseaux ont permis aux opérateurs de maitriser les outils de diagnostic et de déterminer les métriques représentatives de la qualité vocale. Cependant, les méthodes de mesure de la qualité audiovisuelle des services conversationnels ne sont pas encore matures et peu exploitées par les opérateurs de télécommunication.

Le présent travail est centré sur la recherche de métriques représentatives de la perception de la qualité des flux associés aux services de visiophonie et de visioconférence. Ces métriques objectives sont calculées à partir du signal audio et vidéo. Des tests subjectifs sont menés afin de collecter le jugement des utilisateurs du service sur la qualité perçue en fonction de différents niveaux de dégradations. Nous avons étudié l'impact des conditions réseau (perte de paquet, jigue et désynchronisation) sur la QoE d'un appel vidéo. Le principe général est ensuite d'établir une corrélation forte entre les métriques objectives sélectionnées et la qualité perçue telle qu'elle est exprimée par les utilisateurs. Les résultats ont montré que les nouvelles métriques de qualité globale audiovisuelle qui prennent en compte l'aspect temporel de la vidéo sont plus performantes que les métriques basées qualité d'images. D'autre part l'utilisation d'une approche machine learning représente une solution pour générer un modèle de prédiction de la qualité globale à partir des métriques de dégradation (flou, pixellisation, gel d'images, …)

Mots-clefs : Qualité d'expérience, qualité audiovisuelle, service conversationnel, évaluation, mesures subjectives, mesures objectives.

Abstract

In a highly competitive environment, one of the key challenges for operators and providers of video telephony services is to ensure the highest quality of experience (QoE). There is a strong need for a measure that reflects users satisfaction and perception of these services. The audio-visual quality of a video call must be controlled to meet two main needs. The first concerns the planning of new technologies under development. The second is focused on the control of existing communications by assessing the quality of the services offered and evaluating them.

Two approaches are used to evaluate audio-visual quality: subjective tests by collecting scores given by participants on quality scales, after viewing and listening to audiovisual sequences and objective metrics based on automatic audio / video or audiovisual quality evaluation algorithms. Concerning telephony services, decades of research, standardization work and network exploitation, have allowed operators to master the automatic monitoring tools and to determine the representative metrics of voice quality. However, the metrics for measuring the audiovisual quality of a conversational services are not yet mature and not exploited by telecommunication operators.

The present work focuses on finding representative metrics of the perception of the video telephony and videoconferencing services quality. These objective metrics are calculated from the audio and video signals. Subjective tests are conducted to collect the judgment of service users on the perceived quality according to different levels of degradation. We studied the impact of network conditions (packet loss, jitter and desynchronization) on the QoE of a video call. The general principle is then to establish a correlation between the selected objective metrics and the perceived quality as expressed by the users. The results showed that new metrics of overall audiovisual quality that take into account the temporal aspect of video are more powerful than image quality based metrics. On the other hand, the use of a machine learning approach represents a solution to generate a global quality prediction model from the degradation metrics (blur, pixelization, image freezing, ...)

Keywords: Quality of experience, audiovisual quality, conversational service, evaluation, subjective measures, objective metrics.

INSA RENNES

N° d'ordre : 18ISAR 02 / D18 – 02
Institut National des Sciences Appliquées de Rennes
20, Avenue des Buttes de Coësmes - CS 14315 - F-35043 Rennes Cedex
Tél : 02 23 23 82 00 – Fax : 02 23 23 83 96