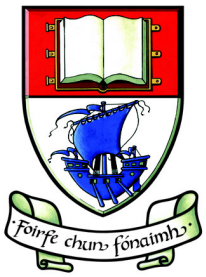
A Framework to Assess Video Quality for Unified Telephony Communications in Enterprise Networks



Himanshu Dadheech, B.Tech

Department of Computing, Mathematics and Physics

Waterford Institute of Technology

Thesis submitted in partial fulfilment of the requirements for the award of

Masters by Research

Supervisor : Dr Brendan Jennings

September 2013

Declaration

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Masters by Research is entirely my own work and has not been taken from the work of others save to the extent that such work has been cited and acknowledged within the text of my work.

Signed:..... ID: 20055555
Date: September 2013

Dedication

My parents

Shashi Sharma & Govind Gopal Sharma

and

to principal source of my inspiration for a research career

Prof V Sinha

Acknowledgements

I would first like to thank my supervisor Dr Brendan Jennings for believing in me, his guidance throughout my research and his support towards independent thinking. I would like to extend my thanks towards Dr David Malone and Jonathan Dunne for being my principal investigator from FAME and research supervisor in IBM.

I would like to thank my parents Shashi Sharma and Govind Gopal Sharma for supporting and encouraging me to work hard and harder and to allow me to pursue my educational career with my choices with their blessings and large-hearted support. My brother Sudhanshu Dadheech for being an internal source of courage and aspiration. My grandparents for bestowing their blessings on me. My uncles and aunts and all other family members for encouraging and supporting me always in all my ventures through out my life.

I am thankful to my research team and friends for supporting my work and helping me in all ways. Specially to Mohamed Adel and Yi Han for being such an enthusiastic team and working together. Personally, I would like to thank all my friends in past and present to help me out and keeping faith in me always and making it possible for me to work far away from home. All my colleagues at TSSG for the great lunch and tea discussions. It has been a great ride along you all and I thank everybody involved in TSSG for putting up so hard work in bringing up the organisation as a family.

I would like to thank the FAME and all the people associated with it for providing me such a great opportunity. Specially, Martin Johnsson, for providing every sort of support for my work. I would also like to thank IBM Software Lab, for providing their facilities and resources to carry forward my research there. Special thanks to Jonathan Dunne for being such a friendly supervisor at IBM and being such a good human helping me out in every possible way.

Abstract

Nowadays there are many audio-visual solutions involving person to person, multi-party telephony and web conferencing, collectively termed Voice/Video over IP (VVoIP). End users expect the quality of VVoIP to be either better than or at least equivalent to traditional telephony. Thus, it is very important for Enterprises to analyze the performance of their systems onto the network with respect to existing technology. Being a totally IP based service, VVoIP is much more susceptible to network fluctuations and end-user application deployment. As an important requirement of end-user quality assurance, this work describes a novel testing framework for assessing and analyzing the call quality of unified telephony in enterprise networks in terms of Quality of Experience (QoE) as perceived by the end user. This framework includes analysis of the system performance under varied network conditions by emulation of network impairments. This framework is suitable for network planning as it formulates the relationship between network and application parameters and objective QoE scores. Packet loss pattern is another issue which plays an important part in video QoE assessment. Video-over-IP packets encapsulate encoded image frames of different type, functionality and size viz. I, P and B frames. Thus, one IP packet dropped could result in losing single or multiple frames having different level of importance and could impact the overall QoE in different ways depending on the particular frame types that were lost. A critical situation for video QoE is when important packets are lost which can deform not only a single frame but consecutive frames as well, reducing QoE from an expected level in the given network conditions and codec implementation. This work additionally presents a study on the impact of loss of specific types of frames on overall video QoE.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction to Video Telephony

The telecommunications industry has taken a large leap over the last few decades in terms of its infrastructure, service capability and quality of service. Public Switched Telephone Networks (PSTN) were implemented many years ago to carry analog voice signals over circuit switched conventional telephone cable Varshney *et al.* (2002). Later digital signal processing made its way into the technology sector and encouraged better use of resources in the telecommunications industry as well. Later Global system for Mobile Communications (GSM) and Universal Mobile Telecommunications Systems (UMTS) came into existence with the use of digital signalling and wireless transmission. The telecommunications industry has seen a rapid change in its technology and implementation in the last two decades. From PSTN to GSM and UMTS and now 3G, 4G and LTE have been implemented for technological advancement and efficient usage of resources with best deliverable quality in civil, defence and enterprise business networks.

Voice over Internet Protocol (VoIP) Goode (2002) has been one of the fastest growing ideas in the field of communications networks. It has laid an infrastructure capable of delivering service to all dimensions of user space with efficient resource management. It has changed the contemporary view of voice signal transmission from one point to another and has established a contemporary infrastructure for voice transmission. Moreover Video and Voice over Internet Protocol (VVoIP) have now been a current trend in research with increasing demand of innovative applications to satisfy the global

needs of multifaceted and quality assured services in the telecommunications industry. The trend of using VVoIP in conferences is increasing day by day in enterprise business sector and researchers have been developing new protocols to efficiently use the present and future infrastructure of the underlying networks.

Video transmission over analog networks has been used as a broadcasting service for television. Nowadays with digital signal processing, analog video signal is first digitized, processed for error concealment, packetized and sent over underlying network to be received at the receiver end and then converted back to analog signal for display. A point of keen interest here is the underlying network, which earlier was a fixed line dedicated wired network like PSTN, having its own capacity and quality delivery issues and now is the IP or the Internet which offers best effort service to the users. Apart from the underlying network, video transmission has taken another leap in terms of variation in types of service demand. Video now is not only used for broadcast for television and other proprietary applications but has a multi-dimensional service demand viz. video telephony, video conferencing, video on demand service, Internet video services, multimedia Internet and mobile applications and IPTV (Internet Protocol TeleVision). Due to these varied service demands, Video over IP has become a significant part of network traffic in nearly all communications networks.

Throughout this technological advancement in the telecommunications industry over the last few decades, the key point of evolution has been to make the network more robust, increase the capacity of communication links, deploy scalable solutions and increase the quality of service being offered. VVoIP has been a crucial concept in finding solutions for all these problems. Its best advantage over any other legacy communications networks is its usage of IP as backbone. Using IP as a service carrier, it incorporates flexible deployment of services, easily scalable solutions, faster and higher capacity data links, easy integration of IP networks with fixed line networks and lower cost of operations and maintenance. Due to these advantages, mobile telephony sector has been attracted to use VVoIP as its operator delivery service base in next generation networks based on all IP infrastructures.

As a part of more user centric and interactive services in telecommunications, video over IP has been promoted as a key solution for being a large part of communications software industry. Large number of companies worldwide are using video chat, video conferencing, and video lecturing services as part of their proprietary telephony service

software. This interactive feature of telephony has already attracted enterprises to migrate to all IP services for their inter- and intra-domain communications. VVoIP has played a crucial role in defining new technology for enterprise networks.

Video over IP in telephony or simply called Video telephony came into existence as a solution for the demand for better user interactive communications and to transmit high amount of data to users. The underlying infrastructure in the 90s for this purpose was huge and the efficiency and capacity of network was not utilized efficiently. With the vast deployment of IP and the Internet, the telecommunications industry decided to move to IP as its backbone as it was more economy friendly for them. Video broadcasters decided the same and then video telephony over IP evolved as a cheap and readily available solution for the industry. With the additional benefit of saving travel expenses and conducting meetings, video telephony has become a key solution for the enterprise environment Kent & Tepper (2005). With highly efficient video codecs which could compress the huge video data to be transmitted by large factors, it was convenient and beneficial to use video telephony. Moreover, real-time delivery of video data in RTP over UDP was also a key feature used as a baseline for global conferencing.

Video coding technology has created a far more robust environment for efficient and secure data transfer over long hauls. Complex video encoding algorithms could compress the raw data by a factor of 80%. Moving Picture Experts Group (MPEG) ISO & IEC (1988), a specialized and dedicated group for development of video encoding standards established by International Organisation for Standardization (ISO) and International Electrotechnical Commission (IEC) have developed some of the best video encoders and decoders in last few years. Starting from MPEG-1 to MPEG-4, it has delivered video coding standards for compression and transmission of video. The basic layout of the video coding standard is in the form of Profiles and Levels which define the set of tools that are available and range of appropriate values for the properties associated with them respectively. Joint Video Team (JVT) was a joint project between International Telecommunications Unions Telecommunications standardization Sector (ITU-T)-Video Coding Experts Group (VCEG) and ISO/IEC-MPEG to develop a high end encoding standard which could be used on varied platforms independent of the underlying or overlying technology and protocols known as MPEG-4 part 10 or H.264 Advanced Video Coding (AVC) in 2001. H.264 AVC Wiegand *et al.* (2003) is now the most used encoding standard in all forms of video services worldwide. Later in

2010, the Joint Collaborative Team on Video Coding (JCT-VC) was established with a group of experts from ITU-T-VCEG and ISO/IEC-MPEG to develop High Efficiency Video Coding (HEVC) which could further reduce the data compression ratio to half in comparison to H.264 AVC, increase the video quality and support higher resolutions. Currently H.264 Scalable Video Encoding (SVC) Schwarz *et al.* (2007) is the encoding standard being used as a solution for scalable encoding and reducing the spatial and temporal video resolution and delivering lower quality video stream as needed by the application. It is a part of H.264 AVC and contains video sub-streams which reduce bandwidth usage and hence better resource management in crisis scenarios.

Moreover, after compression of video data by highly efficient video encoding streams, error concealment algorithms have made it easier to transmit video data securely from one point to other. Error concealment algorithms not only reduce the probability of losing image frames from video but also aid in increasing the capacity of the system and hence the overall network data rate of video transmission. This additional benefit helps in delivering high resolution video streams with low loss. Forward Error Correction (FEC) algorithms using convolution codes help in secure and fast data transmission Nguyen & Zakhor (2002). Integration of high end data concealment schemes with sophisticated encoding standards have supported delivery of high resolution, high performance video streams with good quality of service.

To deliver a quality product in the market there have been keen efforts from industry to measure the performance of the system and make sure that the final product delivers the promised quality as assured in Service Level Agreement (SLA) between the vendor and the client. Performance analysis of communications software has been in the interest of standardization bodies as well viz. ITU, ISO and IEEE. As part of the requirement of quality assurance for clients, standardization institutions have developed standard metrics to measure network performance in terms of Quality of Service (QoS) CISCO (2001). In field of telephony QoS is defined by ITU as the set of requirements for all aspects of communications such as response time, loss ratio, signal to noise ratio, cross talk, echo, interruptions, frequency response and so on. A service or protocol which follows the QoS establishes a traffic contract with application software and reserves the capacity or resources in network nodes and links beforehand for quality assurance. In these terms QoS guarantees the resource allocation and per-

formance for traffic flow but does not guarantee the overall performance of the system. QoS does not suffice all conditions to represent overall performance of a system.

IP networks do not provide dedicated network resource management service to the end users during a call in contrast to legacy systems and contemporary wireless mobile networks in case of voice calls. Apart from that the Internet or IP could assure best effort service and protocol. Video over IP being an all IP based service also could not guarantee affixed network resources required by the end user and hence is highly vulnerable to network fluctuations and value of network parameters like bandwidth, end-to-end delay, jitter, network capacity, packet loss ratio and noise. These network parameters, if not taken care of could degrade the video telephony service to its worst level. Apart from the network resources, application level parameters also play an important role in defining the overall quality of the video phone service Seshadrinathan *et al.* (2010a). As discussed earlier, codecs and error concealment algorithms play an important role in video transmission in any video service. Hence choices like codec implementation, codec profile and parameters level for a particular network scenario could affect the quality of video delivery severely. Apart from codec choice and implementation, optimization algorithms implemented within the communications software client like jitter buffer optimization and forward error correction algorithm could also collaborate in defining the quality of video delivered to the end user. Susceptibility to network and application layer parameters make the quality of video phone service vulnerable and play an important role in defining the SLA with end user.

Dependence of video phone service onto varied number of parameters makes it more important to test, analyse and verify the quality of video delivered to the end user by the application. Moreover being a user interactive service, it is more important to keep the end user perceived quality in mind while delivering the product. Hence it is highly recommended to evaluate and optimize the product for quality as perceived by the end user.

1.2 Motivation and Research Questions

With the increasing interest of enterprise business sector in video telephony applications, it has become an important issue for all software and service vendors to deliver a solution which is:

1. Efficient in resource management in both hardware and network domain in comparison to legacy and contemporary services;
2. Supportive of variable service types within one platform viz. audio, video, conferences, text and slide sharing;
3. Not vulnerable to the sudden fluctuation in the available network resources;
4. Guaranteed solution for quality of service and performs above the benchmark level of industrial standards.

QoS objectively measures the service delivery and most of the time is not related to the end user but the media. For an end user, QoS is not a sufficient performance metric to represent quality of the product Van Moorsel (2001). Quality of Experience (QoE) Le Callet *et al.* (2012), is a subjective methodology to measure the end user experience with the service offered. The need of user perceived quality for the product leads to use of QoE instead of QoS. Standardization institutions have defined a metric to represent QoE for end user as Mean Opinion Score (MOS). MOS measures the user experience with the service and represents it in numerical form in the range of 1-5, where 5 being the best and 1 being the worst. Being a subjective methodology involving end user experience with the product QoE in form of MOS represents a sufficient and necessary measure for evaluation of quality of video over IP services as it encapsulates network, application and even hardware level impairments and degradation issues.

Literature review presented in Section 2.8 highlights the lack of testing and evaluation methodologies and an autonomic framework to compute user perceived quality for VVoIP products. Moreover, there has not been a full grown solution within the enterprises to evaluate and further enhance the quality of their product using the existing industrial standards for QoE computation. User perceived end-to-end quality has not yet been a significant topic of interest in the enterprise world.

Hence motivation of our work comes from the requirement of testing and evaluation methodologies for quality assurance of video over IP communications software and the need to incorporate QoE instead of QoS into the analysis procedure. QoE being a subjective methodology also involves large human interaction and interception with the original product. It requires a group of individuals to sit down and score the actual product under variable scenarios possible. This cumbersome task of involvement of

humans into the testing process is another key motivation for our work as it arouses a need of an automated objective assessment method for estimating QoE of the end product as perceived by the end user. The research question that arises from this requirement is:

- *Can we realise a QoE Assessment Framework from readily available software components that facilitates comprehensive and repeatable studies of VVoIP clients?*

Moreover the dependence of video quality delivered to the end user on codec impairments is also an important issue to be studied. Codec implementation is in the purview of the application developer and has nothing to do with the underlying physical resources which depend on the external environment factors like temperature, location, distance and cost of deployment. Codec is a part of software deployment which depends only on the computational power and space available as resource. Dependence of overall video quality on the encoding scheme could be studied and optimized for better performance even in unhealthy network environment. This motivates us to understand the basic encoding methodology used by H.264 AVC codec and find out the relation between video MOS generated by the service with few coding parameters under varied network scenarios. Even dependence of video quality on the video source is of importance to optimize the performance of the communications software for best achievable QoE. This requirement leads to a research question;

- *How to analyse the performance of a given video over IP product or specific codec in a given network environment in terms of user perceived quality and perform a comparative analysis of different products?*

Moreover network loss is not the only defining parameter to affect the video QoE in a particular manner. Same network loss could result in different different perceived quality for different video contents. Different codecs could perform in a non similar manner for the same network environment. Apart from external environmental conditions and network environment codec implementation is an important factor to be considered for assessing or maximising the overall video QoE. Packet loss pattern, video content, key-frame interval, packet loss at specific locations of video sequence and video codec Group of Pictures (GOP) implementation are some factors to be considered as

important deterrent for video QoE. Dependence of video QoE on these factors lead to a research question;

- *How does specific packet/frame type loss affect the overall video QoE?*

1.3 Contribution

The contribution of this thesis to the scientific and industrial knowledge is in three different ways.

First one is the investigation and implementation of different testing methodologies to test, assess and analysis the performance of communications software in enterprise networks. This thesis investigates the testing methodologies to evaluate objective MOS score for end-to-end performance evaluation of the telephony application. Moreover, it proposes a novel and automated testing framework generic to all telephony applications to run test-cases and evaluate the MOS and analyse the performance of VVoIP based communications software. The proposed framework acts as a plugin to the telephony application to evaluate its performance and produces QoE results in the form of MOS. It uses a suite of external tools to run the testing and evaluation process. Test results are generated for a varied number of network scenarios or environment conditions by emulation of network impairments for the communications links.

Second one is the accuracy analysis of the present standard objective models to predict the MOS based on network statistics only. ITU-T has provided Opinion models for audio and video telephony to estimate the MOS score based on the network impairments without requiring the source and degraded voice or video sequence. This thesis presents a study of how accurate these models are and what other conditions and parameters need to be taken care of by these models for more precise emulation of human behaviour to generate QoE results in the form of MOS. This study also presents MOS from real-time tests in enterprise environment and analysis of performance of some widely used enterprise and civil telephony applications.

Third one is the study of impact of specific frame type loss on the overall QoE of the video telephony application. This thesis presents the functionality of different frame types encoded by H.264 AVC codec and their impact on the QoE. IP packets encapsulate encoded image frames which could be of different type and having different functionality and size. Thus losing one IP packet could result in losing a frame or

multiple frames having different level of importance. One key frame lost could be far more dangerous for QoE than five other frames lost together. These loss patterns make things worse for video QoE when more important packets are lost which can degrade not only a single frame but few consecutive frames as well, bringing the actual QoE level down from the expected level in given network conditions and codec performance. This work focusses on study of the impact of loss of specific types of frames on over all video QoE.

1.3.1 Research Publications

1. Dadheech, H.; Han, Y.; Jennings, B.; Malone, D.; Murphy, L.; Dunne, J.; Sullivan, P.; , A Quality-of-Experience Assessment Framework for Management of Enterprise Voice/Video-over-IP Services , Communications Magazine, IEEE, (submitted and under review) Dadheech *et al.* (2013a)

2. Dadheech, H.; Jennings, B.; Dunne, J.; , A Call Quality Assessment and Analysis Framework for Video Telephony Applications in Enterprise Networks , Global Information Infrastructure and Networking Symposium (GIIS), 2013, (Accepted) Dadheech *et al.* (2013b)

1.4 Thesis Structure

This thesis includes 6 chapters and describes step by step progression of study through each chapter. Chapter 1 presents a brief introduction to Voice over IP and Video over IP technology, the need of measurement of quality of VVoIP systems, introduction of QoE and its advantages over QoS to encapsulate end-user experience, motivation of this work and research contributions. The requirement to deliver quality product to the consumer and to reduce the production cycle time of communications software based on VoIP requires an extensive process to test, evaluate and analyse the performance of these products from the perspective of end user. Although QoS is a major part of SLA between vendor and client, but it is not the sufficient measurement of quality perceived at the end. For the same reason, QoE has been introduced to represent the quality of voice or video as perceived by the end user. Motivation of this research lies in objectifying and automating the QoE assessment process. Chapter 2 introduces the basic technical concepts behind voice and video over IP technologies and the protocols

used in them. It also introduces the external tools used for specific purposes within this research and presents some related work done in this area of research. Chapter 3 introduces an framework to test communications systems inside enterprise environment. It presents different methodologies and their integration into a single framework to analyse the performance of a communications software. It also presents some sample results for few audio and video codecs. Chapter 4 discusses issues related to quality assessment of video telephony and produces results for QoE assessment of video telephony using various popular video telephony services. Chapter 5 introduces the concept of dependence of QoE on packet loss pattern. It describes various frame types and their functionality, encoded by H.264 AVC codec. It presents results showing dependence of video QoE on specific frame loss under different packet loss scenarios. Finally chapter 6 presents conclusions and future prospects of this research.

Chapter 2

State of the art

This chapter covers the details of the technology studied, tools used and related literature reviewed and describes the steps to drive this research. This chapter presents an overview of the VVoIP technology and the formation of problem statement for this research. The technical concepts behind measuring quality of experience and its metric are discussed next. Perceptual Evaluation of Speech Quality (PESQ), opinion model for speech quality estimation also known as E-Model, peak-to-peak signal to noise ratio (PSNR) for video and opinion model for video quality estimation and Video Quality Metric (VQM) are discussed. Furthermore, this chapter describes the voice protocols, codecs and video protocols and codecs used in this research. It mentions the tools used for measuring various application specific and network specific parameters and processing of media data and generation of MOS. Finally it discusses the related studies to this research.

2.1 VoIP

VoIP stands for Voice over Internet Protocol or generally referred as Internet Telephony. VoIP is one way of voice communications similar to PSTN and cellular mobile telephony, all differing in their backbone infrastructure. As the nomenclature itself provides the idea that it is based on the Internet protocol or IP, it uses the Internet as its carrier of media data. Due to low availability of the Internet in most of the public areas, this technology was subjugated in the past. Moreover due to its requirement of fast Internet connection, it was not as popular as other forms of communications in

former years. Recent developments of laying down vast Internet networks worldwide also known as broadband, has enabled use of VoIP in public domain and increased its charm in industry to produce more robust protocols, infrastructure and applications to use VoIP as a business solution Ramakrishnan & Kumar (2008).

VoIP is used for the same purpose as any other communications service viz. PSTN, i.e. transferring of voice data from one point to another or between a group of users in case of multi-party calls, but on a different network infrastructure. VoIP allows its users to use Local Area Networks (LAN) or Wireless LAN ($WLAN$) or Wide Area Networks (WAN) using a reasonable broadband Internet connection as its media carrier for voice data. At the user end, the end equipment could be either an IP phone or a softphone software application that could capture and convert voice signals into digital form to be transported over the Internet link to the other end. The fundamental concept of VoIP is converting analog voice signal into digital signal, packetize it into IP packets and send it over the Internet link to be received on the other side of call, then de-packetized and converted back to analog voice form to be played back on end user equipment. VoIP service can run on privately owned IP networks, such as enterprise networks and leased lines, or the public Internet, which can be accessed via Internet service providers ($ISPs$). VoIP clients include hardware platforms like PCs, PDAs, smart-phones and dedicated VoIP boxes or other communications device with access to the Internet and software application clients that could run on these hardware platforms. The overall VoIP service network consists of physical voice interface, end user client, hardware platform, access network and core network. A general diagram of overall VoIP service is shown in Fig. 2.1.

VoIP basic call procedure could be explained in 4 steps; digitization of speech signal, compression of digitized data, packetization of compressed data and transmission of IP packets via the Internet link. A general flow diagram of VoIP point-to-point call is shown in Fig. 2.2 Salah (2008). Digitization of speech signal is the basic fundamental of signal digitization which includes sampling and quantization of analog signals. Crucial parameters here are sampling rate and quantization intervals which define the precision of reconstruction of speech signal at the other end. Since human voice varies from 300 to 3000 Hertz, it leads to basic voice-frequency transmission channel of bandwidth of 4KHz. According to Nyquist-Shannon sampling theorem, sampling frequency should be at least twice the voice frequency for effective reconstruction of speech signal. Thus,

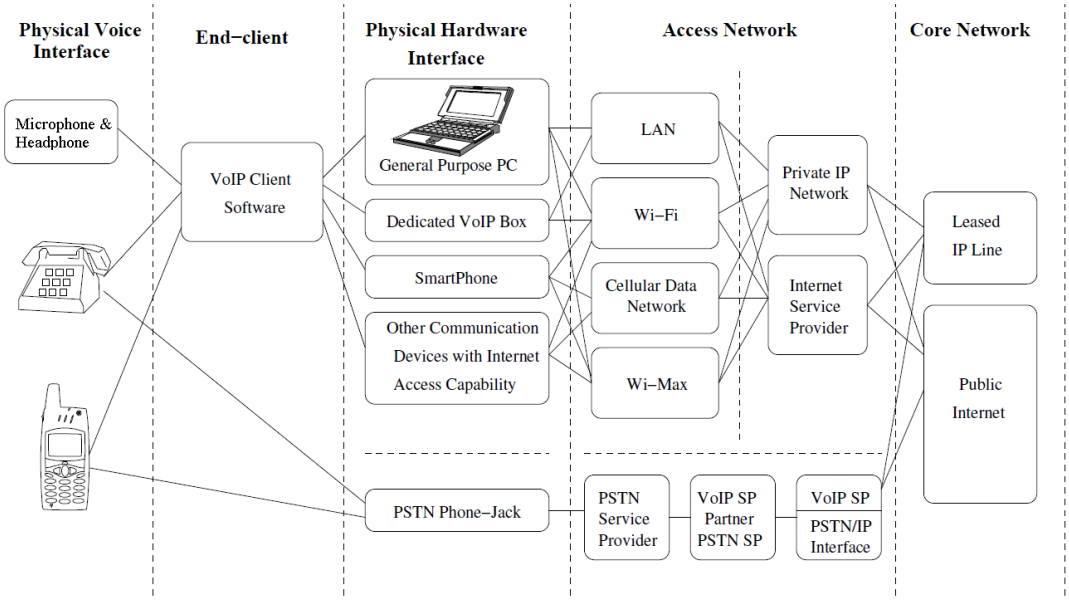


Figure 2.1: General VoIP System Infrastructure

sampling frequency of 8KHz is used in voice digitization. Data is present in binary format after digitization of speech signal. This huge raw data requires high bandwidth in network link which is highly cost consuming for service providers and end users. This, encoding of raw data is done for compression so that redundant data could be dropped while transmission or only that part of data could be selected which is sufficient and necessary for speech signal reconstruction. For example, silence and higher frequencies could be cut down to reduce data set size for actual transmission and few bits are added for the receiver side to know the presence of these obsolete components in the speech signal for accurate reconstruction. This job of smartly coding the media data for transmission is done by codec. Codec may vary in their specific functionality and resource requirement like compression ratio and bandwidth requirement. Speech codecs are discussed in detail in section 2.5 of this chapter. This encoded data is then packetized in the form of IP packets and IP header is added at the start of each packet. At this stage, these packets resemble any other form of IP data packets transmitted using the Internet except the header fields and specifications. These IP packets are then transmitted over to the other end using IP over the Internet and the reverse process is acquired to play back analog speech signal on end user equipment.

Ding *et al.* (2007) quotes that VoIP has emerged as an important application and

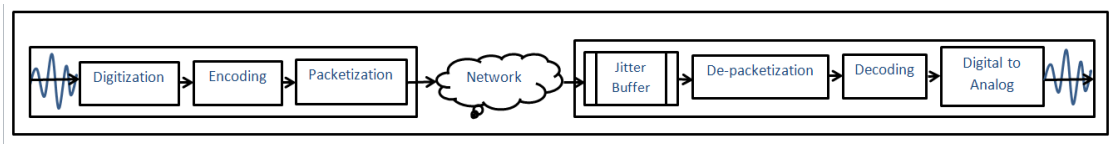


Figure 2.2: End-to-end VoIP call components

it is expected to replace the current Public Switched Telephone Network in next few years. In fact the main reason behind the popularity of VoIP is the reduced cost of deployment of underlying infrastructure and its maintenance and the easy scalability of service for futuristic view. Next section covers video over IP description.

2.2 Video over IP

VVoIP stands for Voice and Video over Internet Protocol and also known as multimedia telephony. Here we will discuss the video telephony part of VVoIP. Video telephony is one of the fastest growing trends in telecommunications industry. With availability of faster broadband Internet connection, video could be streamed in real time over IP and thus could be used as more interactive method of communications as compared to voice telephony. Recent studies Atenas *et al.* (2010), Schierl *et al.* (2009), Kim & Yoon (2008) show the importance and vital necessity of video telephony in computing world. According to these studies, this technology will grow even faster in coming years.

Video over telephony uses the same concepts as VoIP, only differing in media type, packet size and encoding algorithms. Being a technology with media data which is few times larger than voice data traffic, video telephony requires more sophisticated encoding algorithms, error concealment algorithms and more network resources in terms of bandwidth. In VVoIP, voice and video traffic is carried on different channels and the requirement of bandwidth for video could be 1000 times the requirement for voice traffic. Moreover video being a more user interactive technology is more prone to network environment fluctuations and degradation. Having a requirement of high bandwidth, video telephony is generally based on infrastructure having IP as backbone because other contemporary network operator services are unable to dedicate such high network resources to a pair of users. Here also the same procedure is followed for digitization and packetization of data, and the only difference is that the size of IP packets is far larger than in voice telephony. As end user equipment, video telephony could utilize PCs,

Figure 2.3: Sample Video Phone

PDAs and smart-phones with functionality of capturing live images using a camera and proprietary video phones that come with the subscription of VVoIP services. A sample picture of video phone is shown in Fig. 2.3. Video telephony clients running on general purpose hardware like PCs, PDAs and smart-phone utilize the existing interfaces as keyboard, camera and monitor. These soft clients have the advantage of scalability and upgradation over the dedicated hardware clients.

A video telephony client differs from VoIP client in two ways; first the user interfaces are different and second the encoding of media stream is different owing to the nature of raw data. User interfaces here are a camera to capture the live images and a display monitor to play those live images in continuous form which appears as a moving picture. Resolution of human eye and brain interface in time domain governs that if more than 24 images per second are shown on a screen in continuous form, then human eye is not capable of detecting the changing period or dark period in between the images. This concept is used while capturing images using a camera. The camera captures images in bursts with a speed of generating more than 24 frames per second. Frame capture and playback speed also define the quality of video, more frames per second gives clearer motion in video sequence and hence better quality. Images captured by the camera are already in digital format. Each image is divided into pixels, which is the smallest resolved unit of an image. Each pixel is defined by its properties like brightness, contrast and colour component. These properties per pixel are stored in binary format and thus there is no need of digitization of video sequence.

A standard 640x480 pixel image in greyscale requires 2.45 Mbits of memory space, which aggravates to bandwidth requirement of 70Mbps for a video sequence with 30

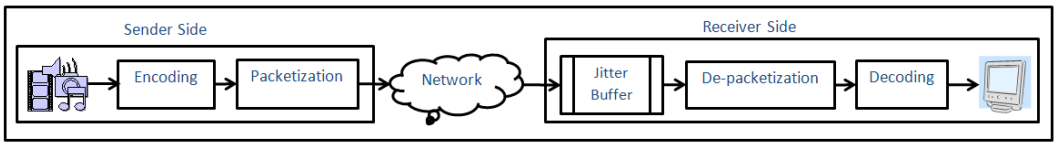


Figure 2.4: End-to-End Video Call Components

frames per second Greengrass *et al.* (2009a). Typical uncompressed video bit rates are 270 Mbps for standard-definition (SD) and 1.485 Gbps for high-definition (HD). These high video bit-rates could restrict the usage of video even on broadband networks and for this reason we need data compression before transmission. This data compression should be done in a way so that the final data size is adequate to be transmitted on normal broadband links and also the reconstruction of video sequence should be lossless. Encoders for video or video codecs are used for this purpose which could compress the image size by even 80%. According to latest encoding standard, each image is divided into portions having nearly same data content or with high correlation of data in its neighbourhood. These portions are known as image slices. Image slices are divided into macro-blocks and macro-blocks are divided into blocks which are groups of 8x8 pixels in the image. Each slice is further classified as I, P or B-slice Kwon *et al.* (2006). Details and functionality of different slice types is discussed in Section 2.7. These slices are encoded in a way that it reduces data content of the overall video sequence. A codec produces stream of image slices which is then packetized into IP packets for transmission and the reverse process of de-packetization and decoding is performed for playback of video sequence on the other end. A data flow diagram of video over IP is shown in Fig. 2.4.

There are different session initialization and transport protocols for video over IP. These protocols differ in their features and affect the quality of video delivered in a different way. These protocols are discussed in section 2.4. Apart from that, there could be different types of video over IP services having different resource requirement and different level of quality delivery. A list of few typical services with their resource requirements is produced in table 2.1 Sims (2007). Futuristic studies show the high demand of video over IP services on all types of electronic communications devices. For catering to these demands more complex and user centric issues need to be addressed viz. quality assurance, enhancing the performance of video over IP, deciding optimized

Appliance	Service	Bandwidth
Television	High Definition TV (MPEG2)	19 Mbps
	Pay TV	3-6 Mbps
	Standard Definition TV (MPEG2)	3.5 Mbps
Personal Computer	Interactive TV on Internet	1-3.5 Mbps
	Video on Demand (VoD)	3-6 Mbps
	Personal Video Recorder	Up to 6 Mbps
	High Speed Internet (WEB Surfing)	Up to 2 Mbps
	Interactive Gaming	1-5 Mbps
	Video on PC	4-12 Mbps
Telephone	Voice over IP (VoIP)	20-64 kbps
	Voice over DSL (VoDSL)	40-64 kbps/chn

Table 2.1: Video services over the Internet and bandwidth required

IP packet lengths depending on the underlying network and robust error concealment algorithms for maintaining video quality at end user equipment.

Moreover, VVoIP performance depends on the network impairments experienced by IP packets during transmission. End-to-end delay, jitter, packet loss, packet loss pattern are some of the key issues that affect the quality of video delivered. We have investigated in this study, the implications of these factors on overall quality of video telephony as experience by the user. Next section would discuss the methods of computing or estimating Quality of Experience (QoE) as perceived by the end user for voice and video telephony.

2.3 Quality of Experience Computation

Although quality of service lays down certain conditions in service level agreement between vendor or service provider and the customer to ensure the quality of the service, yet it does not guarantee a benchmark quality and could not measure the actual quality as perceived by the user. Quality of Experience (QoE) is the real measure of quality of a service or product or application as perceived by the user. QoE is measured in terms of the Mean Opinion Score (MOS).

Being a user experience based metric, MOS in itself is difficult to be measured and standardized Kuipers *et al.* (2010). Traditionally MOS for audio or video is obtained

Quality Scale	Score
Excellent	5
Good	4
Fair	3
Poor	2
Bad	1

Table 2.2: Mapping of human quality perception and MOS

by subjective listening/viewing tests in which a group of users will listen to/view the original and degraded media streams to score the quality of received media. MOS ranges from 1 to 5, where high score indicates better quality. The crude relation of MOS score and human perception is presented in table 2.2. However, we emphasize on the point that scores could vary depending on user group, environment and required service level Brooks & Hestnes (2010). Subjective tests require a large group of human resources and are time consuming and there is no perfect algorithm to score a video by the end users and hence results are not reproducible Jumisko-Pyykkö & Häkkinen (2005). For this reason it is not feasible to perform subjective tests every-time on VVoIP products and hence we have developed an automated framework to do this task during this research. The framework is explained in detail in chapter **3**.

There are two major methodologies for testing video over IP services; intrusive/offline and non-intrusive/online methodologies Moorthy *et al.* (2010), Winkler & Mohandas (2008), Winkler (2009). Intrusive methodology includes comparing the original and degraded sample transmitted through the system and then producing degradation in quality in terms of required metrics viz. MOS. The MOS is designed to range from 0.5 to 4.5 as video and voice over IP systems. A MOS result of 4.5 is considered as the highest achievable score and 0.5 as the lowest. Non-intrusive MASUDA & HAYASHI (2006) method gathers data in form of network and application parameters on the run and predicts the quality in terms of MOS based on its algorithm without even looking into media data. An introduction to most generally used metrics and algorithms and tools is presented in the following subsections.

2.3.1 *PESQ*

PESQ stands for Perceptual Evaluation for Speech Quality and is one of the widely used objective MOS calculating tools. *PESQ* is a standard based on Rec (2001), Rix *et al.* (2001) and comprises a test methodology for objective assessment of voice telephony as perceived by the user.

According to the results presented in Rec (2001), *PESQ* has demonstrated acceptable accuracy for codec evaluation and codec selection decision process in a given network environment. For assessment of voice quality in VoIP systems, sample original audio file could be used as a reference and input into *PESQ* to be compared against its replica received over the other end of VoIP call under evaluation. The *PESQ* algorithm compares the signals in the two samples, calculates the difference and finally gives an evaluation of the quality of degraded sample as an estimation of human perception. The *PESQ* score is mapped from 0.5 to 4.5, but the output range is mainly in between 1.0 to 4.5 which is the normal range of MOS values that were found in listening quality experiment Rec (2001).

For convenience and saving costs on subjective tests, *PESQ* analysis is used instead of subjective test methodology when the experimental set is large. For example, while performing experiments for a new VoIP client in mobile networks for its quality evaluation, large number of tests could be conducted in a controlled manner and over a real or emulated network infrastructure and the input and output media samples could be recorded and passed as input to *PESQ* for MOS generation.

PESQ requires the original sample and degraded sample for comparison and thus cannot be used in real time, simultaneous to active calls. MOS produced by *PESQ* varies from subjective tests as described in Rix (2003). This shows that *PESQ* does not perfectly emulate human behaviour of scoring media sample quality. Moreover, *PESQ* does not take into account the content type i.e. language of the content or dialect, which could introduce some variation in MOS by subjective tests.

MOS listening quality ($MOS-LQ$) was proposed and mainly aimed at giving results that are correlated to subjective tests by applying the 3rd order regression mapping function in 2.1, where x is the MOS from PESQ and y is the corresponding $MOS-LQ$:

$$y = \begin{cases} 1.0, x \leq 1.7 \\ -0.157268x^3 + 1.386609x^2 - 2.504699x \\ +2.023345, x > 1.7 \end{cases} \quad (2.1)$$

The mapping function shifts the MOS results from *PESQ* tool closer to the human perception, which is what E-Model tries to represent as discussed below. In this research, MOS-LQ is used for more accurate comparison with E-Model values as it is closer to subjective test results and is applicable to a wider range of network types (fixed, mobile, VoIP). Next section discusses mathematical model to estimate MOS for audio depending on network statistics and codec used.

2.3.2 E-Model

E-model is a set of mathematical equations to model the human behaviour for quality estimation of voice signals. It depends on network statistics during the call and audio codec used in the call. It does not require any media source or received file for comparison and hence could be used as an objective model for online estimation of voice calls.

The E-Model Rec (2003) is the most popular objective MOS estimation methodology. It is a non-intrusive method that accepts network characteristics during the call and codec information as inputs and outputs an estimated call quality score in real time. The output of E-Model is the "Rating Factor R" which can be mapped to MOS scale. E-Model was standardised in 2005 in Rec (2003) and further extended to wideband codecs in 2011 in Rec (2011).

Fig. 2.5 shows the transmission parameters used as input in the computation model. Room noise of sender person (Ps) and room noise receiver person (Pr) representing environmental background noise and D-Factors represent noise caused by the microphone and loudspeaker, which may vary from sender and receiver side and the values are handled separately in the algorithm. The parameters Sender Loudness Rating (SLR), Receiver Loudness Rating (RLR) and circuit noise (Nc) are referred to 0 dBr point by default. Other parameters including sum of SLR and RLR (Overall Loudness Rating, OLR), Quantizing Distortion (qdu), Equipment impairment (Ie), and advantage factor (A) are considered as values for the overall connection. The other parameters including Side-tone Masking Rating (STMR), Listener Sidetone Rating (LSTR), Weighted Echo

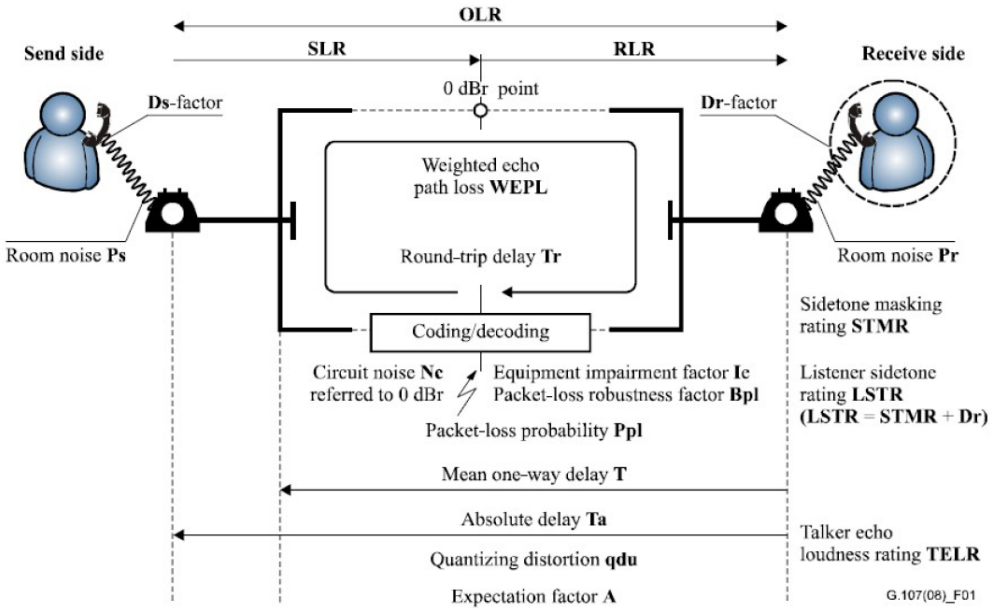


Figure 2.5: E-Model algorithm parameters connections Rec (2003).

Path Loss (WEPL) and Talker Echo Loudness Rating (TELR) are values considered only for the receiver side.

The Rating Factor R combines all transmission parameters for the connection and is calculated by:

$$R = Ro - Is - Id - Ie_{eff} + A \tag{2.2}$$

where Ro represents the basic signal-to-noise ratio, including noise caused by the circuit and background noise; the Is factor is the combination of all impairments that occur more or less simultaneously affecting the voice signal; Id represents the quality degradation caused by delay and Ie_{eff} represents the quality degradation caused by low bit-rate codecs and also includes the degradation due to packet losses; the advantage factor A is an adaptive value that in many cases is constant.

I_d is a function of the one way delay only and can be calculated by the approximated formula expressed in Eq. 2.3 Cole & Rosenbluth (2001):

$$I_d = 0.024xd + 0.11x(d - 177.3)xH(d - 177.3) \tag{2.3}$$

where

$$H(x) = \begin{cases} 0, if x < 0 \\ 1, if x \geq 0 \end{cases} \tag{2.4}$$

Ie_{eff} is function of the codec used and the packet loss rate, it can be expressed by:

$$Ie_{eff} = Ie + (95 - Ie)\frac{P_{pl}}{P_{pl} + B_{pl}} \tag{2.5}$$

Here Ie represents the impairment factor given by codec compression, B_{pl} represents the codec robustness against random losses and P_{pl} represents measured network packet loss rate. The values of Ie and B_{pl} are given only for ITU codecs in ITU-T G.113 appendix ITU-T (1994) as neither the impairment factors of all the codecs factors are provided nor can they be calculated easily. ITU-T recommendation G.113 does not provide codec Ie and B_{pl} values for the most commonly used codecs like $ILBC$, $SILK$, GSM and $SPEEX$.

Language variance leads to a further variance of MOS from subjective testing score when converted from E-Model to R score. A Japanese version of MOS was proposed in Takahashi *et al.* (2006), in which a linear regression mapping function was applied to convert the E-Model MOS to Japanese E-Model MOS. Next section describes the most popular video quality measurement metrics.

2.3.3 $PSNR$

Peak Signal-to-Noise Ratio is the ratio between maximum possible power of the signal and power of noise that corrupts the signal and affects its fidelity. $PSNR$ is the most used measure of quality of reconstruction of lossy compression methods used by video codecs. For images, original image is the signal and noise is the error introduced due to network impairments or codec compression in the received image. $PSNR$ is the approximation of measure of reconstruction quality and ability of images by a codec as perceived by humans. It essentially computes the amount of error present in the degraded image as compared to original one. It ranges from 0 to 50, the higher the better. A separate algorithm has been used to convert $PSNR$ value into MOS and is described in Chapter **3**.

$$PSNR = 20.\log_{10}(MAX_I) - 10.\log_{10}(MSE) \tag{2.6}$$

Typical acceptable value of $PSNR$ for video codecs is 30 to 50.

One drawback of using $PSNR$ as a metric to measure quality of video telephony is that, it cannot be used for real time monitoring of calls as it needs both original and degraded media streams for comparison and scoring. Secondly, we need an algorithm to represent its scores in terms of MOS which in itself needs modelling by subjective testing and training.

One advantage of using $PSNR$ is that it is widely used metric for scoring image compression models and precisely represents the end-to-end error occurred during transmission.

2.3.4 SSIM

Structural Similarity Index (SSIM) Wang & Li (2007), Wang *et al.* (2003a) is a method to compute similarity between two images. It uses a measurement approach based on structural distortion within images. Structure and similarity here refer to signal samples having strong dependence on each other specially when they are close in space. When pixels which are close to each other have strong dependence on each other, their structural forms seem similar and it plays an important part in human perception of objects within the image Wang *et al.* (2004).

SSIM is more perceptual as compared to PSNR or MSE as it computes image degradation as the perceptual difference in the structural forms. The main concept of this metric lies in the fact that the human visionary system is highly skilled in extracting structural information of objects from the viewing field.

2.3.5 VQM

Video quality metric as described in ITU-T G.1070 Rec (2007) is a standard metric for quality measurement and benchmarking of video transmission systems. It measures the quality of transmitted video in a range of 1-5, 5 being the best and 1 the worst. VQM is a non-intrusive/in-line quality estimation tool as it does not require the source data to be present for comparison and scoring. It uses major network and application level parameters for quality estimation viz. packet loss ratio, delay, bandwidth allocated, codec impairment parameter, sending frame-rate and bit-rate. VQM is a result of an integrated function which takes in the effect of codec distortion and network losses.

The basic video quality affected by coding distortion i.e. the effect of video encoding and decoding on the quality of video sequence, I_{coding} is represented as;

$$I_{Coding} = I_{Ofr} exp\left\{-\frac{(ln(Fr_V) - ln(O_{fr}))^2}{2D^2_{Fr_V}}\right\} \tag{2.7}$$

where O_{fr} represents the optimum frame-rate to maximise I_{Coding} for each sending bitrate Br_V and $I_{O_{fr}}$ is the maximum video quality expected at each Br_V with corresponding O_{fr}. D_{fr_V} represents the degree of video codec robustness depending on frame-rate.

$$O_{fr} = v_1 + v_2 Br_V, 1 \leq O_{fr} \leq 30 \tag{2.8}$$

$$I_{O_{fr}} = v_3 - \frac{v_3}{1 + (\frac{Br_V}{v_4})^v_5}, 1 \leq I_{O_{fr}} \leq 4 \tag{2.9}$$

$$D_{Fr_V} = v_6 + v_7 Br_V, 0 < D_{Fr_V} \tag{2.10}$$

$$V_q = 1 + I_{Coding} exp\left\{-\frac{P_{plV}}{D_{P_{plV}}}\right\} \tag{2.11}$$

Other than I_{Coding}, V_q depends on $D_{P_{plV}}$ degree of video quality robustness due to packet loss and P_{plV} represents packet loss ratio.

$$D_{P_{plV}} = v_{10} + v_{11} exp\left\{-\frac{Fr_V}{v_8}\right\} + v_{12} exp\left\{-\frac{Br_V}{v_9}\right\} \tag{2.12}$$

Coefficients v_1, v_2, .., and v_{12} are dependent on codec type, video format, key frame interval, and video display size.

2.4 Voice over IP Protocols

This section discusses the major VoIP signalling and transport protocols in brief. These protocols have their own functionality and features which affect the quality of media delivered in different ways. Some of the well-known signalling and transport protocols which are also used in this research are covered below.

2.4.1 SIP

SIP stands for Session Initiation Protocol and was developed by Internet Engineering Task Force (IETF) and it is an application-layer signalling communications protocol. SIP is used for creating messages transferred between peers for the purpose of creating, modifying and terminating VoIP calls. SIP could be used for different applications such as instant messaging, video conferencing, online gaming and voice communication over the IP. While other VoIP protocols are not capable of supporting TCP, SIP has the ability to redirect calls through UDP and TCP. Various features of SIP include allowing media to be added or removed during the sessions and also permitting multicast conferences in the existing sessions Lambrinos & Kirstein (2007), Fathi *et al.* (2006), Hoeher *et al.* (2007). As quoted in Rosenberg *et al.* (2002) "SIP transparently supports name mapping and redirection services, which supports personal mobility-users can maintain a single externally visible identifier regardless of their network location".

2.4.2 RTP

RTP stands for Real-time Transport Protocol and it defines standard packet format for delivery of audio/video data over IP network. RTP characteristically acts over UDP (User Datagram Protocol) for transfer of data and provides services like multiplexing and checksum. It does not provide resource allocation or assurance of promised QoS, however it expects lower layer protocols to perform these tasks. RTP is used in conjunction with the RTP Control Protocol (RTCP). While RTP carries the media streams (e.g., audio and video), RTCP is used to monitor transmission statistics and QoS and aids the synchronization of multiple streams. As quoted in Horak (2007), "RTP does not either guarantee delivery through the network or prevent out-of-order delivery, and it does not assume that the underlying network is reliable and delivers datagrams in sequence to the receiving machine.

2.5 Voice over IP Codec

ITU-T has proposed and developed a number of codecs for audio compression and decompression. Some of the most commonly used are $G.711$, $G.721$, $G.722$, $G.723$, $G.726$, $G.727$, $G.728$, and $G.729$. These codecs differ in their compression ratio, compression

algorithm, packet size, bandwidth requirement and features. This section presents a short summary of these codecs.

2.5.1 *G*.711

G.711 is used as a standard voice codec in IP telephony, but it requires higher bit rate as compared to other codecs Roychoudhuri *et al.* (2003). Thus, the rate of a call in one direction is 64 Kbps. Furthermore, enhanced features were added into *G*.711 such as *G*.711.1 also known as Wideband Embedded Extension for *G*.711 pulse code modulation. According to ITU Rec (1988), *G*.711.1 contains an annexure that addresses the usage of *G*.711.0 with *G*.711.1. *G*.711.1 spans the bit-rates range of 64, 80 and 96 Kbps.

2.5.2 *G*.723

G.723 was developed to provide real-time coding and suitable voice quality for voice transmission as a modified extension of *G*.721. *G*.723 codec is not suitable for music and provides lower quality output than other codecs Cui *et al.* (1998). According to Menth *et al.* (2009), "it was specially designed for voice encoding at low bandwidth and is mostly used in VoIP applications, e.g., in Netmeeting or Picophone. *G*.723.1 can operate in two different modes generating 6.4 Kbps with 24 byte chunks or 5.3 Kbps with 20 byte chunks every 30 ms".

2.5.3 *G*.729

G.729 uses Conjugate-Structure Algebraic-Code-Excited Linear-Prediction (CS-ACELP) algorithm to compress payload to low bit-rate. *G*.729 provides less delay and high speech quality Ding & Goubran (2003). According to Varga *et al.* (2009), *G*.729.1 codec is the first speech codec with "an embedded scalable structure built as an extension of an already existing standard. It offers full backward bitstream interoperability at 8 Kbps with the much used *G*.729 standard in voice over IP (VoIP) infrastructures". Table 2.3 shows comparison between basic codec implementation parameters of these three codecs. Table 2.3 shows some specifications of audio codecs Muppala *et al.* (2000).

Codec	G.711	G.723.1	G.729
Coding speed (kbps)	64	5.3/6.3	8
Frame Size (ms)	20	30	10
Processing Delay (ms)	20	30	10
Look ahead Delay (ms)	0	7.5	5
DSP MIPS	0.34	16	20
Payload (bytes)	160	20/24	20
Number of Flows	7	84/71	56
Subscribed rate packet time	20	30.2/30.5	20

Table 2.3: VoIP Codec Specifications

2.6 Video Codec

There are two popular video codecs discussed and evaluated in this research H.264 and VP8. Both of these are developed by different organisations addressing different needs of specific applications. However, they have been used for VVoIP calls in different applications and thus are important to be evaluated and compared. A brief introduction and comparison of these two codecs is presented in this section.

2.6.1 *H.264*

Most widely used video compression standard is MPEG-4/H.264 AVC which is an image block based Discrete-Cosine Transform (DCT) and motion-compensation-based codec. This standard was developed under a partnership project named as Joint Video Team (JVT) by ITU-T VCEG and ISO/IEC MPEG. Wide use of H.264 AVC could be realised as it is used in all blu-ray players, streaming Internet sources, Adobe Flash-player, Microsoft Silverlight, Apple IPod and HDTV broadcasts.

"The compression and decompression of information is a small component of the MPEG-4 AVC functionality. During the development of MPEG-4, it became evident that there was a need for the standard to cover the streaming of interactive multimedia content over low bandwidth networks and Internet connections" Puri & Eleftheriadis (1998). MPEG-4 is capable of broadcasting different bit-rates ranging from approximately 10 Kbps to 1.5Mbps. MPEG-4/H.264 AVC codec standard is defined in terms of Profiles and Levels which define the set of tools that are available for encoding and

the set of values of parameters or properties associated with them respectively. Profiles are defined for the specific purposes based on applications demand, usage and resources available and are named as Baseline, Main, Extended and High Profiles. Each profile has its own Levels and different features supported by different set of resources and delivering different performance or quality levels. The basic structure of the encoding algorithm involves image blocks, macro-blocks, slices and frames. Profiles differ in implementations of these different types of slices and frames namely I, P and B. Detailed functionality of these frames is discussed in section 2.7. Selected properties of H.264 AVC codec are shown in table 2.4 Chen *et al.* (2006). Table 2.5 and table 2.6 shows some sample profiles and levels for H.264 AVC codec Sullivan *et al.* (2004).

Different types of codecs were selected for experimentation in this research based on availability in the sample communications software clients under test. Different codecs were tested under varied network conditions for performance evaluation of communications software client.

Standard	**H.264 AVC**
Block Size	16*16 to 4*4
Transform	4*4 integer DCT
Entropy Coding	VLC, CABLC, CABAC
Ref Frame	Multiple (5) Frames
Picture Type	I, P, B, SI, SP
Coding Efficiency	2
Decoder Complexity	2.6
Target Applications	DTV, HD-DVD, Mobile Devices, Web-Conferencing

Table 2.4: Video Codec H.264 AVC Specification

2.6.2 *VP8*

VP8 is a successor of VP7 in the VPx series developed by *On2 Technologies* and now under *Google* as a part of *WebM* project Bankoski (2011). Its major target areas were the web applications requiring high quality open video compression format including *HTML5* video. VP8 support high compression and low computational complexity for decoding and some of its features make it different and comparable to some most

Standard	Baseline	Main	Extended
I and P Slices	X	X	X
CAVLC	X	X	X
CABAC		X	
B Slices		X	X
Interlaced Coding (PicAFF, MBAFF)		X	X
Enh. Error Resil. (FMO, ASO, RS)	X		X
Further Enh. Error Resil (DP)			X
SP and SI Slices			X

Table 2.5: Profiles in H.264 AVC Standard

Level Number	Typical Picture Size	Typical Frame Rate	Maximum compressed bit rate (for VCL) in Non-FREXt profiles	Maximum number of reference frames for typical picture size
1	QCIF	15	64 kbps	4
2	CIF	30	2 Mbps	6
3	SD	30/25	10 Mbps	5
4	HD (720p or 1080i)	60p / 30i	20 Mbps	4
5	2kx1k	72	135 Mbps	5

Table 2.6: Levels in H.264 AVC Standard

popular video compression algorithms Bankoski *et al.* (2011). Some of its distinctive features are;

- VP8 adds on lot of features to its predecessor such as Golden Frames, processor-adaptive real-time encoding and low-complexity loop filter and many more to achieve good video quality at low bit-rates
- VP8 was specifically designed to provide video service under quality range of 'watchable video' to 'visually lossless' and hence the requirement of bandwidth is lower as compared to other high quality supportive video codecs
- VP8 supports majority of image formats supported by web video applications

- VP8 uses a different referencing system for inter prediction involving 3 reference frames as compared to other standard codecs
- VP8 uses extensive inter and intra frame prediction. "TM_PRED" mode is used for intra prediction and a flexible "SPLITMV" mode used for inter prediction of images and frames
- VP8 exactly specifies the number of reconstructed pixels and thus validating the implementation of decoder.
- VP8 offers both Variable Bitrate (VBR) and Constant Bitrate (CBR) encoding options
- VP8 was designed for a variety of hardware types ranging from 60MHz processor to multi-core processors. It encodes in real-time on low-end machines and takes quite less cycles to decode as compared to other decompressing algorithms
- VP8 uses high performance sub-pixel interpolation for motion compensation which could give up to one-eighth pixel accurate motion vector.

Moreover VP8 is now compared as competitor to H.264 as being used by *Skype* sky (2013) and *Google Hangouts* gha (2013). Some of the competitive features of VP8 with respect to H.264 are;

- H.264 has three intra prediction modes to estimate contents of a block without referring to other frames whereas VP8 has 4 intra prediction modes
- H.264 uses I, P and B frame system for inter prediction whereas VP8 uses only I and P frames and it has an alternate prediction frame , called Golden frame
- H.264 has built in macro-block quantizer for adaptive quantization whereas VP8 does not
- VP8 loop filter has a wider range while filtering between macro-blocks as compared to H.264.

Apart from the features and latest developments in VP8, it is gaining popularity as a video codec to be used for web-based video services and is considered as a major competitor to H.264 AVC in this research.

2.7 Image Slicing in Video Codec

Motion Picture Experts Group (MPEG) has produced several standards for IP-based services. MPEG-2, VC-1, MPEG-4 Part 10 and H.264 are some most popular encoding standards for video transmission services. With these kind of encoding, the sending data rates have reduced so low that they demand only 3 Mbps for SD and 9 Mbps for HD in case of MPEG-4 Part 10 compression scheme. This encoding or compression helps not only in reliable and faster transportation of video sequence via the network but also adds on capacity to add more security and encryption and additional control channels on the same link.

These compression techniques divide the video sequence into individual images, then part of images called as slices. Slices are further divided into macro-blocks and blocks which add one more level of complexity to the system. Following subsections discuss in detail the formation of image slices and their referencing system for decoding video sequence.

2.7.1 Slice Type and their Functionality

As per the guidelines of MPEG, MPEG-4 Part 10 and H.264 have common structure of image framing and compression. Both of them use different encoding and decoding schemes but the frame types and referencing methodology is same. There are three major frame types I (Intra-coded picture), P (predictive coded picture) and B (bidirectional predictive coded picture). A defined sequence of these frames is known as Group of Pictures (GOP). The size and structure of GOP is specific to codec implementation. Coding of images, frame types and their encapsulation in NAL (Network Abstraction Layer) units and definition of NAL header is discussed in Wenger *et al.* (2005).

Each H.264 frame contains block, macro-block, and slice information. A block is an 8x8 matrix of pixels or their Discrete Cosine Transform (DCT). A macro-block contains several blocks that contain information of a section of frame's brightness and colour component. A slice is a series of macro-blocks and the number of macro-blocks may vary for different slices. Finally a frame contains a number of slices or a single slice. In this work, frame and slice are used as analogous and interchangeable terms here onwards. Every slice in the frame could be encoded as I, P or B. Fig. 2.6 shows a sample image fragmentation into slices, macro-blocks and blocks.

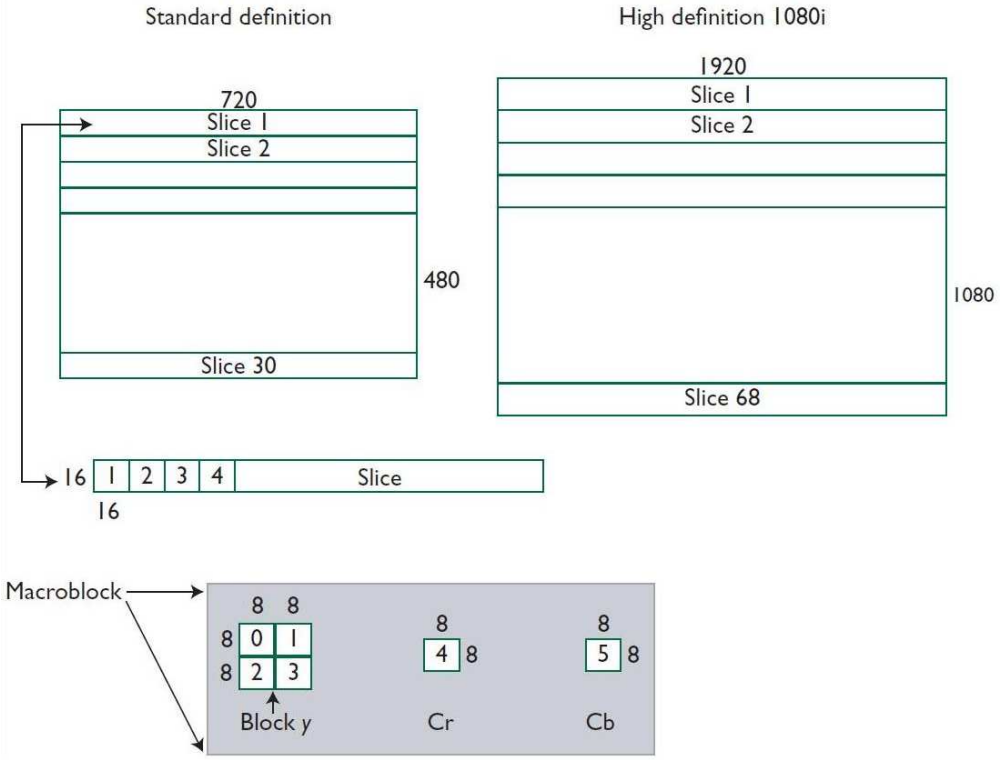


Figure 2.6: Fragmentation of an image into slices, macro-blocks and blocks. Block contains 8x8 chunk of pixels representing colour and brightness information, a macro-block contains a few blocks and a series of blocks form a slice which then accumulates to form an entire image Greengrass *et al.* (2009a)

I-frames or intra-coded frames use spatial compression i.e. it uses the property that neighbouring pixels are correlated to each other. It does not use temporal compression i.e. it does not refer to any other frame for reconstruction of its own image. Spatial compression reduces the size of the actual frame. This infers that I-frames are implicitly encoded and they contain all the necessary information needed for their reconstruction.

P-frames or predictive-coded frames use spatial as well as temporal compression. Preceding I-frames provide reference to P-frames for their reconstruction in the decoder. In other terms, it stores only changes with respect to the preceding I-frame. For example, if an object is moving in front of a stationary background, P-frame will store only the movement of the object in encoded format. This temporal compression further reduces the size of frame to 20% – 70% of the associated I-frame.

B-frames or bi-directionally predictive-coded frames use the previous and the next

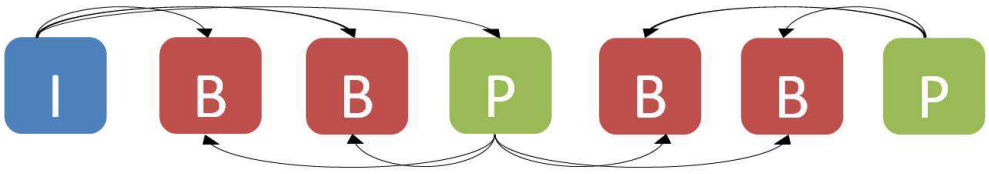


Figure 2.7: Slice reference relationship within GOP. I-slice being the key source of temporal information and thus showing different level of importance for each slice type. Outward arrow represents frame being referred and inward arrow represents dependent frame.

I or P frames to reconstruct their original frame. This bi-directional temporal compression further reduces the size of B-frame to 5% – 40% of the associated I-frame. Here the distance of a B-frame from its reference frames could be an important parameter while considering reconstruction of an image in error prone channel transmission. Some levels of codecs also allows use of B-frames as reference for other B-frames in proximity.

Fig. 2.7 shows a typical implementation of referencing system in between image slices for reconstruction of a whole video sequence. Here the B-frames are not used as reference for any other frames. Dependence of frames on each other and the importance of each frame is visible in Fig. 2.7.

These frames are arranged into a typical sequence for best compression and a set of these frames is called as Group of Pictures (GOP). Structure of GOP is the repetitive unit of the whole video sequence. A GOP contains I-frame and then few B-frames and then P-frame and this pattern continues till next I-frame. As an example, a typical GOP contains 25 frames with one I-frame at the start then two B-frames followed by one P-frame and the pattern of B and P continues till we have the next I-frame i.e. the 26th frame. Size of GOP and the filling of B-frames between I and P-frames depends on the implementation of codec. GOP structure 25:2 means that, it contains 25 frames with two B-frames in between I and P or P and P. A sample 15:2 GOP is shown in Fig. 2.8.

Figure 2.8: Sample GOP structure showing a 15:2 GOP. Every new GOP would contain the same frame sequence Greengrass *et al.* (2009a).

2.7.2 Network Abstraction Layer

H.264 is designed as a technical solution for varied video services like television broadcasting, storage of video on optical or magnetic devices, video-on-demand and multimedia applications etc. These myriad applications demand a common representation of data for all underlying networks and platforms. For this flexibility and customizability, Network Abstraction Layer (NAL) is used which through its header information represents all kinds of encoded video data in a manner which is convenient to a variety of transport layers and networks.

Encoded data is packetized into NAL units. A NAL unit is a self-contained-independently decodable data set. NAL packetization is shown in Fig. 2.9. NAL header contains information in form of the syntax error for H.264 stating whether this NAL unit is to be used to reconstruct pictures using inter picture prediction and payload type. NAL header structure is shown in Fig. 2.10. NAL header plays a crucial role in defining the decoder parameters as the decoder assumes that NAL units are in sequence. NAL header also provides information about payload types which defines the NAL type and is used in our work for identifying the lost frames.

Fig. 2.10 represents the structured format of a NAL header which is used for identifying the type of packet, its features and its payload. **F** stands for forbidden_zero_bit. Functionality of this bit is to indicate the possibility of bit error and syntax violations inside the payload or NAL unit type octate. **F** bit set advices the decoder of the possibility of errors inside payload or NAL unit type header. In this scenario, decoder may decide to discard the NAL unit and conceal data from other NAL units or try to recover the best possible data from the erroneous NAL unit. **NRI** represents nal_ref_idc which is a 2 digit binary code to represent relative importance of NAL unit payload

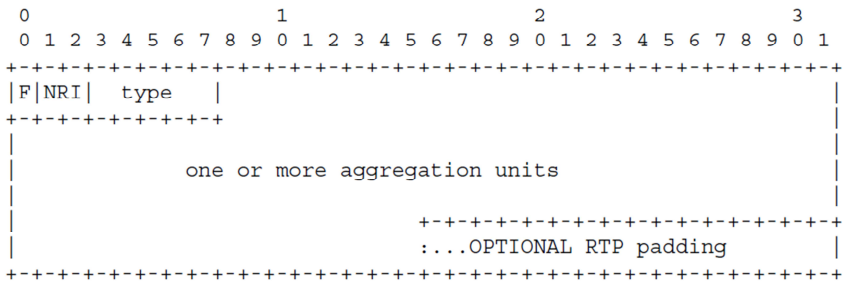


Figure 2.9: RTP Payload format for a single NAL unit containing single or multiple aggregated packets Wenger *et al.* (2005)

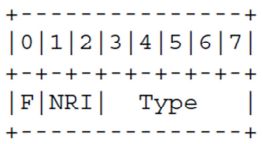


Figure 2.10: Format of NAL Unit type octate Wenger *et al.* (2005)

and presence of decodable payload inside NAL unit. **NRI** value 00 represents that there is content inside the NAL unit to be used for reconstruction of reference picture for inter-picture prediction. The values of **NRI** above 00 represents the relative transport priority and higher the value, more the importance of decoding the NAL unit to preserve the integrity of reference pictures. **Type** specifies the NAL unit payload type. Table 2.7 shows the different payload types for H.264 Rec (2005).

All NAL units containing nal_ref_idc greater than 00 contain data to be used for reconstruction of images. Depending on the **Type**, payload type could be coded data or some other parameter set. For the NAL units with coded data, each NAL unit could contains several slices. Each slice has a slice header and data. The slice header contains information about slice type, type of macro-blocks in slice and number of the slice frame. Type of slice is defined by 2 digit binary variable **slice_type**. Table 2.8 shows the 2 digit binary values of **slice_type** and their corresponding translation into slice types. **slice_type** has been used as an identifying field for the type of slices lost within this research and is included in Chapter 5.

Next section presents a detailed review of the present state of the art literature and

Type	Definition
0	Undefined
1	Slice layer without partitioning non IDR
2	Slice data partition A layer
3	Slice data partition B layer
4	Slice data partition C layer
5	Slice layer without partitioning IDR
6	Additional information (SEI)
7	Sequence parameter set
8	Picture parameter set
9	Access unit delimiter
10	End of sequence
11	End of stream
12	Filler data
13...23	Reserved
24...31	Undefined

Table 2.7: NAL unit types

the research problem formulation.

slice_type	Description or Name of Slice Type
0	P-slice. Consists of P-macroblocks (each macro block is predicted using one reference frame) and / or I-macroblocks
1	B-slice. Consists of B-macroblocks (each macroblock is predicted using one or two reference frames) and / or I-macroblocks.
2	I-slice. Contains only I-macroblocks. Each macroblock is predicted from previously coded blocks of the same slice.
3	SP-slice. Consists of P and / or I-macroblocks and lets you switch between encoded streams.
4	SI-slice. It consists of a special type of SI-macroblocks and lets you switch between encoded streams.
5	P-slice.
6	B-slice.
7	I-slice.
8	SP-slice.
9	SI-slice.

Table 2.8: Name Association of slice_type

2.8 Literature Review

Real time voice and video traffic contributes a good part of data flowing through IP networks. This traffic is also sensitive to network impairments as well as its own application layer parameters. Variation of these network and application layer parameters viz. bandwidth, end-to-end delay, jitter, codec selection, content of media, frame-rate and others have a significant impact on the quality of media and service as perceived by the end user. The study of quality standards in terms of metrics and parameters which are as close as possible to real human interaction with these systems has been an open topic of research worldwide.

This section presents an overview of the related published literature referred during this research. Section 2.8.1 covers the introduction and scope of quality assessment method and metrics for VVoIP applications. Section 2.8.2 discusses work and literature related to methods and metrics required for computation or estimation of QoE. Section 2.8.3 introduces previous work in field of QoE assessment of video applications. It also covers the different testing methodologies and mathematical models involved in

video QoE assessment. Section 2.8.4 discusses some experiments and effects of different external parameters like network loss, bit-rate and codes specific parameters on overall video QoE. Section 2.8.5 discusses some work highlighting the effects of packet loss pattern, loss of key frame, loss duration and frequency of loss on the overall video QoE. Lastly, Section 2.8.6 discusses previous work on effect of video content on the video quality delivered and some amendments proposed in current opinion model for video telephony. A summary of the literature review and problem formulation is described in Section 2.8.7.

2.8.1 Quality Assessment of VVoIP Applications

Necessity of developing a perceptual quality metric for assessment of interactive audio-visual IP based applications has been a keen topic of interest for researchers as well as industry; we now present some related literature in this field.

Keepence (1999) and Savolaine (2001) discuss the definition of QoS for VoIP, its features, limitations, parameters affecting QoS and issues and technologies needed for delivering Voice over IP. Takahashi *et al.* (2004) discuss the idea of quality design and maintenance in VoIP and infers need of a perceptual quality analysis. Moreover, they present a method for perceptual QoS evaluation. Van Moorsel (2001) explains the purpose of QoE and defines the need of QoE where QoS is not a sufficient parameter to encapsulate the perceptual quality of audio-visual services over IP.

All the above presented work emphasises on the functionality and purpose of QoS and also define QoS and QoE. Some of them have attempted to define QoE for perceptual evaluation of VoIP. Neither of them addresses the issue of how to use QoS to determine QoE. In terms of objectively defining QoE metrics, QoS could play an important role in specifying the underlying infrastructure to estimate the performance of VoIP service. However there have been some work done on correlating QoS and QoE and some mathematical models have been proposed in literature to identify the relation between QoS and QoE as presented below.

Fiedler *et al.* (2010) present a model to establish quantitative relationship between QoE and QoS. They propose a generic formula to connect QoE and QoS parameters using an exponential relationship for better QoE control mechanism. Kim & Yoon (2008) proposes a QoE-QoS correlation model to evaluate QoE from QoS parameters and some analysed values from QoE-QoS correlation analysis results for the purpose of

QoE control in IPTV. This model provides a live QoE monitoring method for prompt response to service degradation in IPTV. Agboma & Liotta (2008) proposes a statistical modelling technique to correlate QoE with QoS parameters and to define the degree of influence of different QoS parameters on user perception. This study helps network providers or Internet Service Providers (ISPs) to perceptually analyse the requirement of resources and manage the network resources for better end-user experience.

While these models and services try to define the relationship between QoS and QoE, they fail to emphasize on the need of an objective model and framework to validate the findings. They also ignore the impact of factors other than QoS parameters in defining QoE while assessing performance of any system or service. They also lack in providing a strategy for precise evaluation and analysis of QoE and its corresponding metrics.

2.8.2 Quality of Experience: Metrics and Assessment Techniques

QoE metrics and assessment techniques have long been a topic of interest for the research community and a large body of literature has been published; we now provide a brief overview of this literature.

Pinson & Wolf (2003) presents a study of features, pros and cons of different subjective video quality assessment methods recommended by Rec (2009). These subjective tests are the baseline for performance evaluation of video phone services, modelling of opinion model for objective quality testing and benchmarking performance of video codecs. De Simone *et al.* (2009) provide a wide study and database containing subjective assessment scores for video streaming with H.264 AVC codec. They have made available the uncompressed original and processed video files and their subjective quality scores. Kuipers *et al.* (2010) have discussed in details the quality of experience from viewers point of view and various techniques and tools available for measuring QoE for voice and video telephony. Various other research works have discussed the need of perceptual quality metrics to assess and enhance the performance of audio-visual IP services and have discussed various methods and metrics to encapsulate user perception for quality assessment You *et al.* (2010), Wang (2006), Winkler (2009), Seshadrinathan *et al.* (2010a).

Studies and research mentioned above represent a set of well defined voice and video perceptual assessment metrics. They also mention the need of QoE as a perceptual

method for more accurate user experience. They lack in defining a methodology to assess the system and compute the metrics. A need of an exclusive methodology for QoE could be realised by the fact that it involves user experience and thus more human interaction than just machine computation. They also do not address the issues of subjective versus objective metrics computation and usage.

da Silva *et al.* (2008) emphasise the combined effects of various network and application parameters on conversational quality of VoIP. Through their experiments and investigations, they have discussed the degree of impact of various parameters on conversational quality. Lot of emphasis has also been made upon necessity of replacing subjective quality assessment methods with objective quality assessment methods and objective metrics. Recent studies and surveys have paid importance to developing objective metrics for QoE assessment and discussed different audio-visual objective quality metrics Wang *et al.* (2003b), Koumaras *et al.* (2005), Rix *et al.* (2006).

While the mentioned work emphasise on using objective metric instead of subjective metric for the purpose of removing human error in computation, they fail to define a system to do so accurately. These publications provide an overview of some objective metrics and do not define a generic metric to be used by all users. Moreover, no comparative analysis of different metrics and their specific target functionalities has been discussed in detail.

2.8.3 QoE Assessment for Video Applications over IP

Assessment of QoE for video applications have been a major topic within the quality assessment researchers community; a brief overview of published literature is presented here.

Winkler & Mohandas (2008) and Engelke & Zepernick (2007) have discussed various subjective and objective perceptual video quality metrics and their use. They infer the need of a reduced-reference or no-reference based, reliable perceptual objective quality assessment metric for video applications. Staelens *et al.* (2010) present a methodology of subjective video testing for Video on Demand (VoD) service, enabling same environment and conditions as realised by actual user. Takahashi *et al.* (2008) present an overview of state of the art objective audio and video quality assessment methods and their standardization in ITU. Seshadrinathan *et al.* (2010b) present a study of subjective algorithms to compute QoE and also present a database of reference and distorted

videos and their corresponding QoE scores. Lu *et al.* (2003) have proposed an adaptive resource requirement and usage algorithm based on quality of video delivered in video applications. They have discussed the joint impact of packet loss and encoding rate on video quality and proposed a feedback loop to the application for quality-based adaptation of approach.

The above mentioned literature covers in detail the metrics and methodology of assessing subjective perceptual VVoIP quality. They lack in defining a generalised framework for assessing all these metrics. They also do not present a clear differentiation between different metrics and their mode of usage for a specific purpose. Presented next are some experimentation and implementation work that define a test-bed based on objective analysis of VVoIP call quality.

Nemethova *et al.* (2006) present an assessment study of video streaming application based on PSNR. They have focussed on relation between subjective MOS and PSNR for time variant video signals. Venkataraman & Chatterjee (2011) present a no-reference kernel based module for inferring MOS in real-time video applications. This work computes the QoE for network video applications depending on various parameters without requirement of a reference signal. Chen *et al.* (2009) propose a framework *Oneclick* to capture users' perception of quality of IP based video service. This framework captures users' dissatisfaction as a combined effect of various factors as a single *click* of a dedicated key instead of MOS. Reibman *et al.* (2004) present a framework for monitoring live IP packets to detect the impairments that affect the video quality. This work presents an in-line video quality degradation monitoring method. Cherif *et al.* (2012) have proposed a new perceived speech quality estimation tool based on the Random Neural Network (RNN) approach to derive a non-linear relationship between network impairment parameters and QoE. This model *A_PSQA* is verified for two codecs, *iLBC* and *Speex* against industry specified standard PESQ.

While the presented literature is application specific, it does not define any generalised framework and methodology for video call quality analysis. This research is motivated by the fact of developing generalised approach of quality assessment which is not addressed in a structured way by previous literature.

2.8.4 Effect of External Parameters on Video Quality

Effect of various network and application parameters on video transmission quality and assessment of reliable perceptual quality has been a topic of interest among researchers; a brief overview of literature related to this field is presented.

Calyam *et al.* (2007) introduce a QoE estimation mathematical model for VVoIP calls based on the GAP-Model. They used subjective tests to benchmark QoE and then develop a non-linear model for QoE estimation involving bandwidth, jitter, delay and packet loss. De Rango *et al.* (2006) provide comparisons between subjective and objective methods and metrics for quality benchmarking and assessment of VoIP systems. They discuss the intrusive and non-intrusive methods of objective assessment and outline scenarios of their usage. Conway (2002) suggests a passive method of measuring speech quality in live VoIP calls.

The literature discussed above mention the subjective and objective methodology of testing VVoIP applications and also mention the objective metrics to be computed. They also discuss the difference in two methodologies of testing but do not specify the usage criteria depending on user functionality demand and tools available.

Joskowicz *et al.* (2011) outline an enhancement to ITU-T G.1070 "Opinion Model for Video-Telephony Applications" for better estimation on perceptual MOS using non-intrusive methods. They have included the effect of video content on overall QoE assessment in terms of different spatial and temporal activities within a video. Zinner *et al.* (2010) present the idea of managing QoE in real-time for video streaming. They make use of the video codec H.264 extension for scalable video coding (SVC) to manage the QoE by adaptation of video parameters like resolution, image quality and frame rate. Tasaka & Misaki (2009) identify factors affecting the QoE of audio-visual communications systems in bandwidth guaranteed networks. They have studied two types of tasks, one audio dominant and other video dominant, on QoE and observed optimized pairs of encoding bit-rate and playout buffer time to maximise overall QoE. Tasaka *et al.* (2008) introduce a video delivery method to maximise QoE. They propose a QoE based Switching between error Concealment and frame Skipping (SCS) which decides a threshold of slice errors within a frame to execute either error-concealment algorithms or skipping the frame for the purpose of achieving high QoE. Yamada *et al.* (2007)

propose a *No-Reference* video quality estimation method. They estimate video quality on the basis of the number of macro-blocks with errors unable to be concealed by error concealment scheme. This work provides a no-reference video quality estimation method with intrusive testing.

While the literature presented above mentions a structured framework to be used for quality assessment, they do not generalise the framework. The enhancements proposed for G.1070 do not include the packet loss pattern, which this research finds to be of high importance in defining the overall video QoE. Moreover, they propose quality evaluation models based on no-reference approach, while this research has used full-reference approach for precise and real-time quality assessment.

2.8.5 Effect of Packet Loss Pattern on Video Quality

Packet loss ratio and patterns have been considered as a major influential parameter of video perceived quality. Due to different functionality of frame/slice types, they affect video QoE in a different manner. Being a new area of research interest, very few researchers have attempted to look into specific frame types, their functionality and impact on QoE, the Group of Pictures (GOP) structure and extracting information from Network Abstraction Layer (NAL) header for better estimation of perceived QoE; a brief overview of published literature is present here.

Greengrass *et al.* (2009a) introduce the impact of network factors on video quality of experience and describe the service level requirements for video telephony and streaming services. They have discussed the architecture of the MPEG encoding algorithm for video and introduced different types of image slices and their functionality in defining video QoE. Greengrass *et al.* (2009b) have discussed the impact of packet loss durations on viewer's quality of experience. They have introduced different types of visible impairments in video and discussed the impact of loss of specific type of frames on overall video QoE. They have produced results for packet loss durations and specific slice type loss and their overall impact on video quality. Mu *et al.* (2009) have produced a study of individual packet loss on video quality. They have studied the major influencing factors of video QoE and their impact on overall quality. They have investigated various video artefacts through user experiments and modelled them into visible artefacts for objectifying the process of quality estimation. They have produced results for joint effects of different visible artefacts on perceived video quality. Dai & Lehnert

(2010) have presented results on impact of single packet loss on video services over IP networks. They have analysed the effect of single I-frame loss, its frequency of loss and time distance between individual frame loss on video QoE. They have produced their analysis on individual I-frame loss, short period losses, degree of influence of distance between loss on QoE and threshold frame loss distance to reduce severe degradation in video quality. They infer from their tests that loss distance impacts video QoE more than loss frequency.

Above presented literature has comprehensively discussed the image fragmentation within a codec and the types of slices/frames present in MPEG and H.264 codecs. They have also mentioned the crucial role of specific slice loss in deciding overall video QoE. However, they have not produced any comparative analysis on impact of different slice loss on the QoE. Major attention is paid on I-frame loss and its frequency of loss and its impact on QoE, but cumulative slice loss has not been discussed.

Hohlfeld *et al.* (2008) present a mathematical view of packet loss patterns and their impact on QoE. They have presented a second order statistics for distribution of packet loss over multiple time frames. They have used a Markovian packet loss pattern and 2-state Gilbert-Elliott model for fitting wide range of expected packet loss patterns. These models are used as packet loss pattern simulators to generate appropriate network traffic and study the impact of specific packet loss pattern on QoE in real-time Internet services. Pérez *et al.* (2011) have produced a study on the impact of packet loss on video quality as a function of part of frame affected by packet loss. They have studied the effect of error propagation from one frame to other and inferred that a single crucial frame loss could propagate to a number of consecutive frames and degrade the video quality more than a burst packet loss. Finally, they have proposed a model to make use of network level information in commercially available enterprise products for better resource planning, video encoding implementation and error concealment or retransmission algorithm optimization for the purpose of maximising video QoE. Zinner *et al.* (2010) have discussed the impact of codec implementation parameters as frame-rate, scaling method and resolution on video quality. They have presented results on impact of these individual factors on QoE and proposed a QoE control mechanism by exploiting the codec implementation parameters.

While the above mentioned articles discuss the effect of packet loss pattern and the effect of packet loss on the overall video QoE, they do not discuss the heterogeneous

packet loss scenario involving multiple types of frames lost. Their discussion on different types of artefacts and their role in defining subjective results of video QoE testing has helped the present research in defining tasks for QoE testing. However, they do not focus on the types of frames lost and their individual and cumulative results on video QoE. This research focusses more on extracting information about the type and amount of slices lost during a call and try mapping it to it's impact on the MOS obtained for the call.

2.8.6 Effect of Video Content on Video Quality

Below are some recent developments in the field of identifying and analysing the impact of video content and its spatial and temporal movement elements. They also propose some changes in the present Opinion model for video telephony based on the research findings on the impact of video content on overall QoE.

Joskowicz & Ardao (2009) and Joskowicz *et al.* (2011) have produced studies on short-comings of the present ITU-T G.1070 standard in modelling the perceptual behaviour to estimate video quality metric. Joskowicz & Ardao (2009) produce a study on impact of subjective movement content in video stream on its QoE. Joskowicz *et al.* (2011) propose enhancement to G.1070 and present new values for its coefficients to include the effect of temporal and spatial variance in video into VQM estimation. They have presented a study and the results for video clips with different contents and spatial and temporal activity and their corresponding QoE at no packet loss. Through these experiments, they have derived new values of coefficients for G.1070 model to introduce effect of video content in VQM estimation. Yamagishi & Hayashi (2008) propose a network level parametric model for monitoring video QoE based on network characteristics. They infer that their model is suitable for network planning and monitoring for video streaming services over IP. Yammine *et al.* (2010) propose a method to analyse GOP structure of an encoded stream without accessing the bitstream. They have used noise estimation in decoded video stream as a tool to analyse the periodic behaviour noise variance across I, P and B-frame types. This pattern helps in extracting the GOP structure and period. This work presents an idea to estimate the GOP structure which could be used to inject specific packet loss pattern and study the behaviour of the individual as well as joint effect of different frame type losses.

2.8.7 Summary

In contrast to all the above mentioned literature, this work is focussed on generalising the methodology and strategy for call quality analysis of voice and video telephony systems based on IP. Presented work focusses on the design of architecture and development of a sample software test-bed for an automated testing of voice and video over IP client. At the centre of the research is the proposal of a generalised methodology for perceptual call quality analysis of video telephony. It also focusses on deriving an analytical process for performance analysis of individual video telephony clients and to propose the framework as a tool to compare the performance of different clients and codecs. As an addition, it addresses the issue of impact of specific slice loss on overall video quality. It addresses the impact of individual and cumulative slice loss pattern on the video QoE.

Chapter 3

Testing Framework

3.1 Introduction

Enterprises have embraced Voice/Video-over-IP (VVoIP) technologies due to the benefits of having a single, integrated IP-based communications infrastructure and the attendant lower cost of operations and maintenance. However, management of the VVoIP systems on an end-to-end basis to maintain adequate Quality-of-Experience (QoE) as perceived by end-users, remains challenging. This is particularly the case for geographically dispersed enterprises and for enterprises where employees are mobile and need to communicate via the public Internet. Researchers and industry practitioners have developed a number of quality metrics and techniques for quality assessment of both voice and video applications. However, from a practical perspective, it is not clear how best to integrate these metrics and techniques into a quality assessment framework that facilitates automation of performance testing under varying network conditions.

The Main Objectives of this chapter are;

1. Study of methodologies for objective QoE assessment which does not involve human interactions and produces results as close as possible to human experience;

2. Study of testing methodologies to test the performance of VVoIP application software;

3. Development of an automated framework for objective QoE assessment and analysis of performance of VVoIP communications software in enterprise networks;

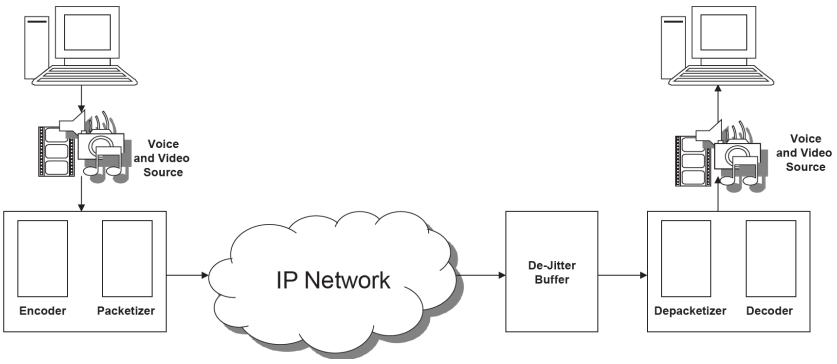


Figure 3.1: General VVoIP architecture: voice/video streams are encoded and packetized at the source, transferred across the network, where they are buffered, depacketized and decoded at the receiver.

4. Propose a sample implementation of the framework using a suite of available tools.

VVoIP performance depends on a number of network related parameters, including: bandwidth allocation, end-to-end delay, packet loss and jitter. Variances in these parameters often leads to degradation of VVoIP applicaiton performance and the QoE perceived by end users Takahashi *et al.* (2004). Moreover, other than these fixed network issues, application specific parameters like the choice of codec, codec parameters Zinner *et al.* (2010) depending on specific network conditions, level of service offered, and jitter buffer sizing also impact on QoE. Given this, it is very important for vendors of enterprise VVoIP applications to test, assess, evaluate and analyse QoE as perceived by the end user. Mean Opinion Score (MOS) is the key metric to measure the QoE of a call as perceived directly by the end user—it encapsulates the effects of both network and application implementation specific issues.

This chapter presents a generalised framework to test and evaluate the performance of VVoIP applications (as depicted in Fig. 3.2) in enterprise networks. The framework could be used to, *inter alia*, assess the performance of an existing deployment, or to assist in network planning before deploying a new system onto available network resources. This chapter describes how the framework can be realised by a suite of open source tools and utilities and presents some results relating to the performance of a sample VVoIP application under a range of network conditions. The chapter is structured as follows: Section 3.2 introduces our Quality Assessment Framework and

its essential components. Section 3.3 describes the suite of tools used for the purpose of implementation of presented framework. Section 3.4 describes how the framework could be implemented via a suite of open source utilities and tools. Section 3.5 describes typical experiments performed using the framework to assess voice and video QoE. Finally, Section 3.6 presents the summary of the chapter.

3.2 QoE Assessment Framework

Our QoE Assessment framework meets the following main functional requirements:

- It should support probing of audio/video data at various points in order to provide input for QoE metric calculation. This is important for the purpose of data collection in different forms and then comparing the results collected at each point in order to analyse the bottleneck of the whole encoding and transmission process;
- It should support a range of QoE assessment techniques, both intrusive and non-intrusive. This requirement enables testing with different techniques to validate the end results, by comparing results collected from different techniques;
- It should provide the ability to emulate a wide range of typical network conditions. This functional requirement enables testing in varied plausible network environments and helps in network planning;
- It should support test case automation—QoE assessment tests should be scripted automatically and be repeatable. Thus, the framework could be employed in the quality assurance processes for VVoIP applications. Automation of the testing process saves time, human effort, reduces human error probability and generalises the testing process for all applications.

The Quality Assessment framework as depicted in Fig. 3.2 is capable of initiating point-to-point audio and video calls and automatically receiving the calls and recording the call data, network statistics and other data related to the call. Configuration parameters to the framework are used to initialize the codec type and its parameters for the given call and the source media to be sent during the call. The synchronization of media during a call is handled, with the recorded media on both sides of the call being

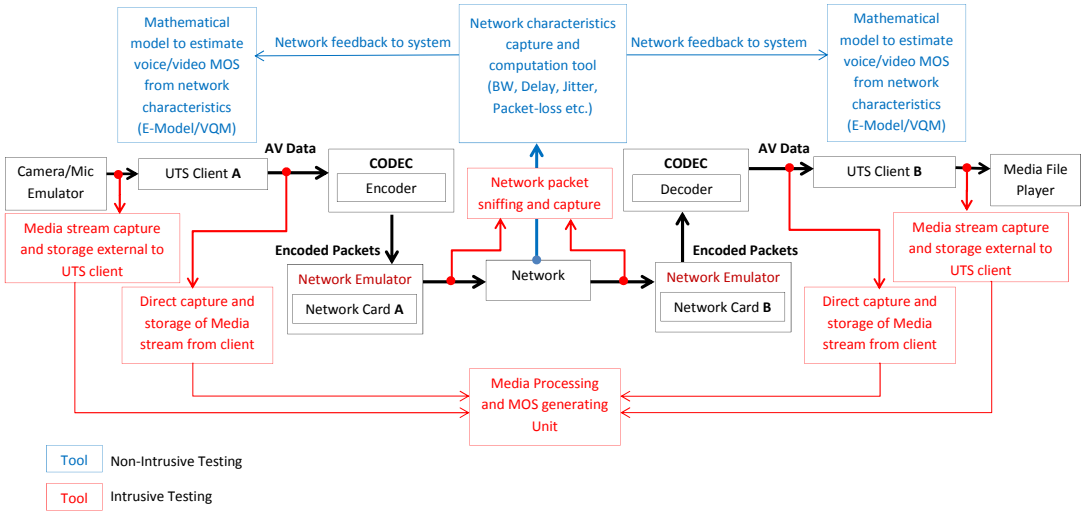


Figure 3.2: Structure of the Quality Assessment Framework, showing: the VVoIP system and under test (black), non-intrusive QoE assessment (blue) and intrusive QoE assessment (red).

properly synchronized in time and frames for both audio and video calls. Variations in network statistics are automated within the framework and can be manipulated to vary on per call basis or even simultaneous to the call.

The framework is designed as a node and probe based network. Each node has specific functionality and takes as input either raw or encoded data and outputs its encapsulation or encoded version or other vital statistics of the call. Since we consider unidirectional flow of data, each link represents different data types having different layer of abstraction. All the probes are placed on the links only so as to collect the data without processing it in-line. The probes then collect data into different nodes having the functionality of processing it and analysing it for producing quality parameters and other results. During this whole transmission, the flow of data through the VVoIP application components is not disturbed.

In Fig. 3.2, the black nodes represent the components of the VVoIP client, red nodes represent the intrusive (off-line) testing workflow and blue nodes represent the non-intrusive (on-line) testing workflow. Starting from the sender side is the camera/mic emulator to emulate the source or input media fed into the software client.

This component is necessary for standardizing the input for reproduction of results under same network environment and application initialization. This node transmits standard media as .wav or .avi to the client. Next component is the VVoIP client user interface which is the gateway for input and output of media data for the client. This component is the significant point for determining the overall quality of the client. Media data in raw form is then transferred to the media framework of the VVoIP client. The media framework of a client is responsible for formatting the raw data in a form which is suitable for transmission; the major component of this media framework is the encoder. The encoder, be it video or audio, is a data processing unit comprising of a set of algorithms collectively called as a standard codec. A codec compresses the media data to reduce its size and defines the final bit-rate, frame-rate, key-frame interval and other intrinsic features of encoded data. Here onwards data is in binary format which could only be read by the corresponding decoder. This data is then encapsulated into IP packets and corresponding headers are added to each packet for transmission over the internet links. For this purpose, the packets have to pass through the integrated network cards of the end-user hardware equipment. The network card defines the identity of the end system and thus the link between the two end points. Network card is connected to the outside network via the internet which is used as the medium for data transfer.

The framework uses a network emulation tool to inject network impairments into the VVoIP traffic. This tool could be configured to set the network scenario as required by the test condition to test the client under varied network parameter combinations. Via the Internet, IP packets reach the other end defined by the network card properties of the receiver. The network card receives IP packets and sends them to the corresponding software application client for processing. Here the client maintains a jitter buffer to take care of variations in delay and continuous playback of media stream. IP packets are then de-packetized, error concealment schemes are deployed (e.g. forward error correction, if supported) and sent to the decoder. The decoder reconstructs the original media stream from encoded data depending on the loss occurred and errors encountered during transmission. Decoded data is sent from media framework to the VVoIP client's user interaction interface for playback via speakers or display on monitor screen using inbuilt required media player. This whole flow of data provides various data sniffing or capture points for the purpose of quality assessment and analysis.

There are three major probing points where we need to sniff data for real time analysis. The first one is direct input and output from speakers for audio and the display screen from video. The probe for this could be seen in Fig. 3.2 on the link from source file node to the VVoIP client node. This could be done using an external client shown as "Media stream capture and storage" node, which could record the speakers or display screen on both sides and then store them for further off-line processing and analysis. The benefit of this method is that it presents the real image or voice as seen or heard by the end user, resulting in a perceptual evaluation of service. The drawback of this approach is that we need off-line synchronization for the media sequences as well as off-line processing like end-points cutting, resolution setting, display brightness and contrast setting.

The second point of data sniffing is at the VVoIP client recording the raw media bits without encoding on both sides of the call. This could be seen in the architecture diagram as the probe on the link between VVoIP client GUI and encoder/decoder. This could be done by recording the media stream when the client receives it from the input camera device on the sender side and outputs it to the display GUI on the receiver side. This functionality is shown in the node "Data capture of media stream from client." An advantage of this probe is that we do not need any off-line synchronization or resolution rearrangement of media streams as it records only the packets of the source being sent and received in its standard format internal to the application client and also the recorded media is independent of the end-users external device settings. A drawback of this approach is that we need a separate plugin to the VVoIP client to set the recording parameters and functionality.

The third probe is used to collect network specific data and parameters during transmission of the encoded stream by the VVoIP clients. The probe is set on the link between the external network and network card of the endpoint equipments. This functionality is performed by node "Network characteristics capture and computational tool." The advantage of this approach is that we get the real time information of network behavior on the media streams, including end-to-end delay, bandwidth allocated, jitter, actual packets lost, type of packets lost for every call placed via framework. The drawback of this approach is that we have to use an external tool to examine the network link outside the network card of the machine.

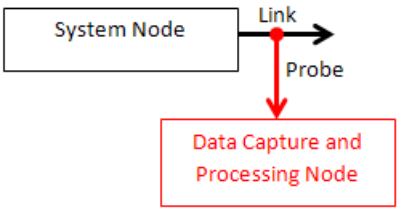


Figure 3.3: Node, Link and Probe structure

After these data processing nodes and probes, we store the collected data into our database for external processing and analysis. For media data collected via probe 1, we use external tools for file cutting, resolution setting, external noise removal and synchronization. After that media data collected from probe 1 and 2 are on the same level of abstraction and could be sent for comparison and generating intrusive metrics for each test-call using standard industry specified tools. This functionality is performed by node "Media processing and MOS generating unit" using a number of tools for data processing and MOS computation. Network statistics collected from probe 3 are fed into E-model or VQM model for audio and video respectively as shown in nodes "Mathematical model to to estimate MOS," to generate MOS scores using non-intrusive/in-line method. These MOS scores are also verified against the MOS generated from probe 1 and 2 to confirm the proper functioning of our testing framework.

3.3 Tools

This section describes the tools used within the research and their functionality.

3.3.1 Dummynet

Testing of a VoIP software client under varied network conditions could be done in operational networks or in simulated environments. While it is not easy to control different operational parameters such as bandwidth, delay, packet loss in the former approach, this research work uses a network emulator. Network simulators are easy to control but they are only approximate models for the desired network settings. This research work uses a real time operational network; however the traffic itself passes through an emulated firewall which shapes the desired network characteristics.

Dummynet Luigi (1999) is a widely used link emulator used to run experiments in user configured environment. It is a simple and flexible network emulator which is capable of delivering accurate network settings with minimal modifications to an existing protocol stack. It could allow experiments to run as a stand-alone system. Dummynet works by intercepting processing of the protocol layer under test and simulating the effects of finite queues, bandwidth limitations, and communications delay and packet loss ratios. It facilitates the features of a traffic generator and protocol implementations while solving the problem of simulating extreme network conditions. Dummynet Carbone & Rizzo (2010) is actually composed of a kernel-level emulator engine, dummynet and its associate packet classifier called ipfw. Both the emulator engine and packet classifier provide large set of features. An ipfw ruleset is made of a list of rules numbered from 1 to 65535 and packets are passed to ipfw and compared against each rule in the ruleset. And these rules trigger the corresponding actions on the packets.

There are two dummynet objects to inject packets into, pipe and queue. A pipe is a link simulator and has its own bandwidth restriction, propagation delay, queue size and packet loss ratio. A pipe operates in First In First Out (FIFO) service model. A queue assigns weight to a particular packet flow. These queues use Worst-case Fair Weighted Fair Queuing (WF2Q+) policy to determine the bandwidth share for each flow.

A pipe could be defined to have a hard bandwidth restriction, or packet loss ratio for incoming packet flow and a queue could further define how the traffic would share the bandwidth. Any other traffic which is not caught under the rules defined for pipe and queue keep flowing through rest of the ipfw without being effected by dummynet. To define its basic operation ipfw feeds packets into a Dummynet pipe or queue and then these packets are buffered and then fed into the pipe depending on the settings of the pipe or queue.

Below is a simple example of ipfw rule to define a pipe using dummynet. The first line adds a pipe for the traffic flowing from one IP address to the other and the next line configures the bandwidth and packet loss ratio for the pipe.

Ipfw add pipe 1 ip from 9.161.46.214 to 9.161.46.238
Ipfw pipe 1 config bw 10Mbps plr 0.08

In this research dummynet has been used as a network emulator to produce network impairments for VVoIP traffic flowing between VVoIP software clients in enterprise environments. The research uses dummynet for emulating delay (latency), bandwidth constraints, and packet loss ratio for outgoing traffic. Next section discusses the tool used for tracing the IP packets from the network for the purpose of collecting live network statistics.

3.3.2 Wireshark and Tshark

Wireshark Combs *et al.* (2007) is a free open source network packet analyser. It is generally used for the purpose of network troubleshooting, analysis, communications protocol development and analysis. Wireshark can capture packets from an incoming stream using *pcap*. It works similar to tcpdump and could also define rules to capture packets from specific source or destination address, specific sequence numbers, protocols and more filtering options. This research work uses Wireshark to capture incoming and outgoing VoIP RTP stream specific to the client IP addresses involved in the testing.

Tshark is the terminal based version of Wireshark which could define the same packet filters as in Wireshark. This research has used Tshark for the purpose of automation of testing process. Below is a sample command to extract the RTP streams associated with each call and some vital information from the captured .pcap file for the purpose of quality comparison and analysis.

tshark -r capture.pcap -z rtp,streams -d udp.port==20834,rtp -q

3.3.3 BVQM and CVQM

Video Quality Metric (VQM) Wolf & Pinson (2002) is developed by National Telecommunications and Information Administration (NTIA) for the purpose of objective measurement of video quality as perceived by the user. BVQM stands for Batch Video Quality Metric, which is a tool developed by Institute of Telecommunications Sciences (ITS) for providing standardized and non-standardized methods for measuring video quality of digital video systems McFarland *et al.* (2007). BVQM is a Windows program for performing intrusive/offline testing. BVQM can perform processing and analysis of multiple video scenes and multiple video systems at once. BVQM reads video sequences from files, and reports results to the screen. BVQM includes a variety of calibration options, quality models, and graphical presentation of results.

Calibration options available in BVQM and used in this research are Full-Reference, Reduced-Reference and No-Reference. Full-Reference compares the original and processed video and calibrates the processed videos frames, the spatial and temporal characteristics and the colour component in correspondence to the original video Wolf (2009). Reduced-Reference uses the reduced reference video calibration algorithm Pinson & Wolf (2005) and No-Reference uses no reference video calibration algorithms developed by National Telecommunications and Information Administration (NTIA). Moreover, quality models present in BVQM and used in this research are PSNR and VQM.

CVQM stands for Command Line Video Quality Metric and is a command line tool performing similar functions as BVQM. CVQM cannot compare results from multiple video clips. This research has used CVQM for the purpose of automation of objective measurement of video quality. A sample CVQM command is show below:

cvqm 'VideoTest1_original.avi' 'VideoTest1_processed.avi' 'progressive' 'frcal' 'psnr'

3.3.4 VDub

Virtual Dub (VDub) Lee (2007) is a free open source video capture and processing utility for Windows. This research has used VDub for the purpose of post processing of captured video sequences. Post-processing includes synchronization of sent and received video sequences in time domain, edge cutting, contrast and brightness matching and resolution matching. Requirements of post-processing are the prerequisites of the video quality measurement tool BVQM.

VDub is also available as a command line tool and it has been used for the purpose of automation of whole testing process. A sample command is shown below:

vdub.exe /s"Processing_Settings.vcf" /b"Original_Video.avi",
"Processed_Video.avi"
vdub.exe /r

3.3.5 ManyCam

ManyCam ManyCam (2013) is a licensed virtual camera emulator utility for Windows. It can act as a virtual camera stream to be fed into different software applications simultaneously which could not be done using only stand-alone camera of machine itself. An additional feature of ManyCam is that it can play a standard video file as a source for the camera video stream fed into different interactive video applications. ManyCam has been used in this research as a virtual camera in sample VVoIP communication client as a video source to send a standard video file in each test-call under different environment conditions. This feature was necessary for standardizing the test input and reproduction of results under same network conditions. A screen shot of ManyCam is shown below.

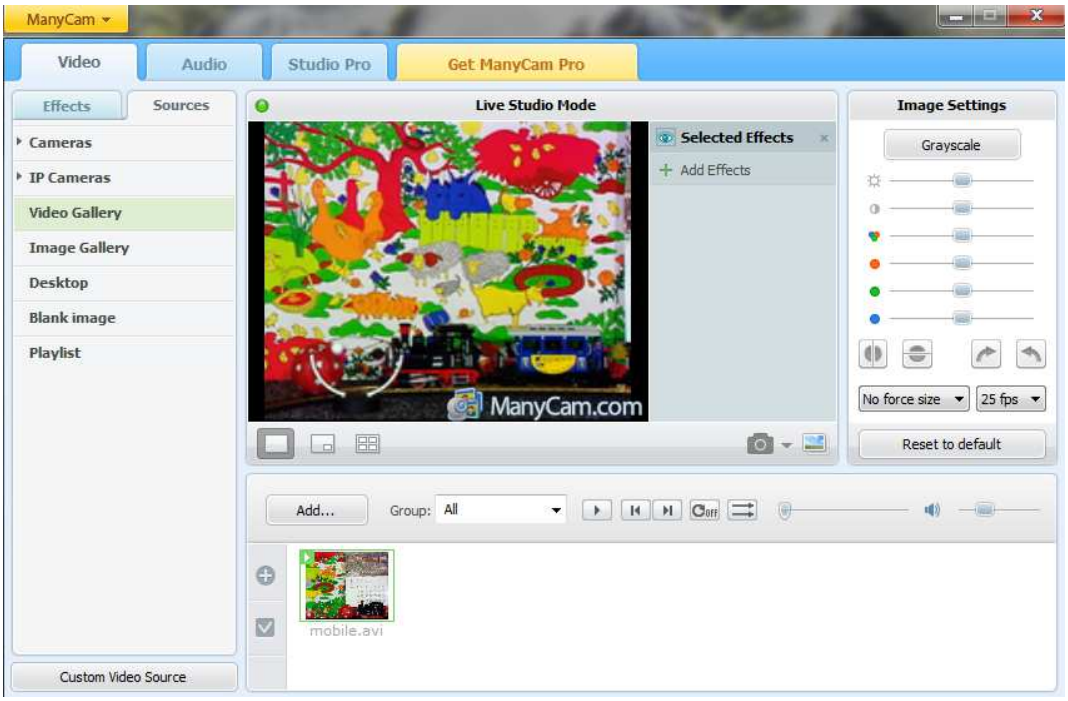


Figure 3.4: Screen-shot of ManyCam playing the source file to be sent using ManyCam as a virtual camera device.

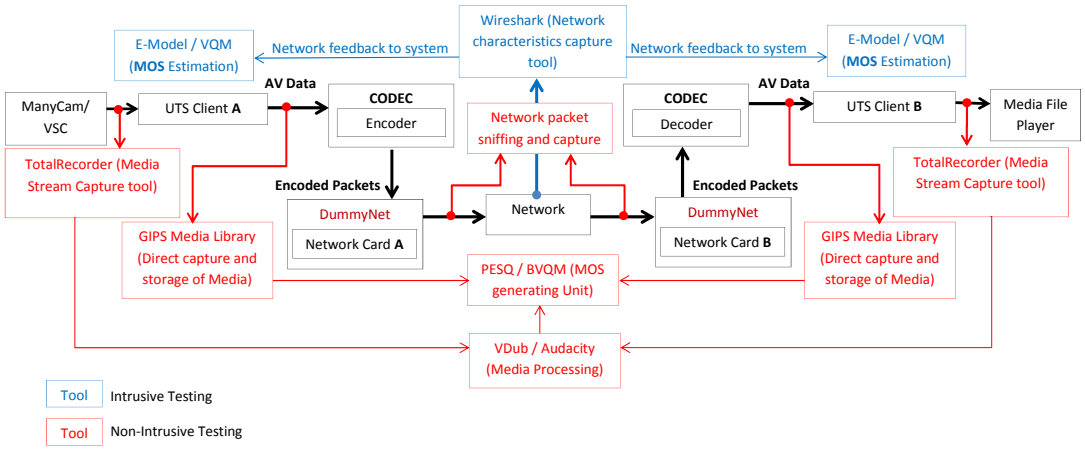


Figure 3.5: Implementation of the Quality Assessment Framework, showing the various tools and utilities used in its realisation.

3.4 Framework Implementation

Our implementation consists of a set of tools and VVoIP client on the two end-points, each having the embedded testing framework as a plugin. Either client on the two sides can have settings to act as a sender or receiver. We have performed a wide range of tests for audio as well as video depending on the combinations of variable parameters for both cases.

The first thing needed in the implementation is a microphone emulator and a camera emulator that accept .wav audio file and .avi video file as input respectively. These emulators are software that could play a pre-recorded standard file as if it is played by real-time user interaction with microphone and camera. In this way, audio/video input can be controlled to be exactly the same for each set of tests and also it is able to bypass the operating system and device impairments that may cause additional degradation of the call quality. Here we have used ManyCam and E2ESoft virtual sound-card for video and audio capture device emulation respectively.

As intrusive measurements are content dependent De Rango *et al.* (2006), Joskowicz *et al.* (2011) and non-intrusive are network and codec dependent Rec (2007), Rec

(2003), the set of parameters used for comparison are: sample contents (language, male or female voice, speech content, degree of temporal and spatial variations in video sequences) and packet loss rate. The audio and video test samples and methods are provided by Recommendation ITU-T P.501, P.920 and P.910 as test signals for use in telephonometry and test samples for the experiments in order to eliminate the risk of choosing inappropriate samples that lead to inaccurate results. The samples selected from ITU-T P.501 are in different languages and have content with both male and female voices. For video, two different video sequences were chosen having different spacial and temporal properties vid (2013). Thus, contents and network variance can be tested under the designed experiment that keeps one variable fixed and changes the other one.

DummyNet Luigi (1999) was used to emulate network and generate packet loss, where for non-congestion related drops, probabilistic match option is used to emulate links with uniform random loss patterns. This ensures that packet loss is not content dependent and could happen at any frame or packet of media sequence. This ensures the randomness of network and relates to real-time traffic emulation. DummyNet results in a normal distribution of packet loss ratio, centred at the requested loss rate value. And thus even if the specified value of packet loss, e.g. 5% for an audio/video call containing 600 packets, such that the expected number of packets lost should be 30, it can in practice vary from 25 to 35, resulting in packet loss to be 4.2%-5.8%. In some extreme cases, the deviation may be larger. Thus the theoretical value of packet loss set in the tool is not used for computing call quality rather the real packet loss rate is calculated by analysing the audio RTP stream.

TotalRecorder was used as media capture tool directly from system's sound-card or screen. Wireshark was used for IP packets capturing and computing various network parameters for each call. VirtualDub and Audacity are used for post-processing of degraded video and audio files respectively. These tools are used for end-point cutting and frame shifting for time and frame synchronization, resolution and brightness setting of degraded files in correspondence to the reference file. Signal power attenuation could be done for audio files for suppressing external noise.

PESQ and BVQM are used for quality metric generation for audio and video respectively. PESQ takes as input the reference file, the degraded file and the sampling frequency and outputs MOS scores. BVQM also takes the reference and degraded video

files as well as the calibration mode and quality analysis mode as input. Full-reference, Reduced-reference and No-reference are available as calibration modes with spatial and temporal shifting parameter configurations. PSNR and VQM are available as quality analysis modes for score generation.

Dummynet, VirtualDub, Wireshark, BVQM and PESQ are available as command line tools and hence available for automation process using some scripting and programming languages. *JAVA (J2EE)* has been used as the programming language to run command-line arguments in a sequential order using the package *java.lang.Runtime*. This package in *Java* enables running the command-line scripts of all the tools mentioned in a programmable order to automate the whole process. Fig. 3.5 depicts how the QoE assessment framework was realised using the tools discussed.

3.5 Framework Usage

In this section we describe a set of experiments that illustrate a typical usage scenario for our Quality Assessment Framework. This section includes results for audio quality in a VVoIP client under test. IBM Sametime 8.5.2 IFR was used as the VVoIP client over IBM lab environment with 100Gbps network connection. For audio, 21 sets of tests were conducted and in each set of tests, network packet loss rate was set from 0% to 8% with 0.5% increments. This overall process is reproduced for each one of the codecs (ITU-T G.729, G.711 μ-law and G.711 a-law). The experimental results are expressed as the MOS values of the point-to-point VVoIP call test-cases against varying packet loss and source content. PESQ and E-Model were used as tools for intrusive and non-intrusive testing methodologies and for their validation. Video QoE results and analysis are covered in chapter 4 in detail.

Fig. 3.6, 3.7 and 3.8 show audio call QoE statistics. Since the loss rates are randomly distributed from 0% to 10%, we have divided packet loss ratio axis into bins of width 0.5% and centred at increments of 0.5%; only average loss rates for each group are presented. Within each bin, there are a number of data points and each data point here represents the mean and median of those points inside the bin. The mean value gives an average score for all the MOS values in its range, including outliers that have extreme low/high values caused by concentrated lost packets in active voice period. On the other hand, the median value gives a more comprehensive value representing the

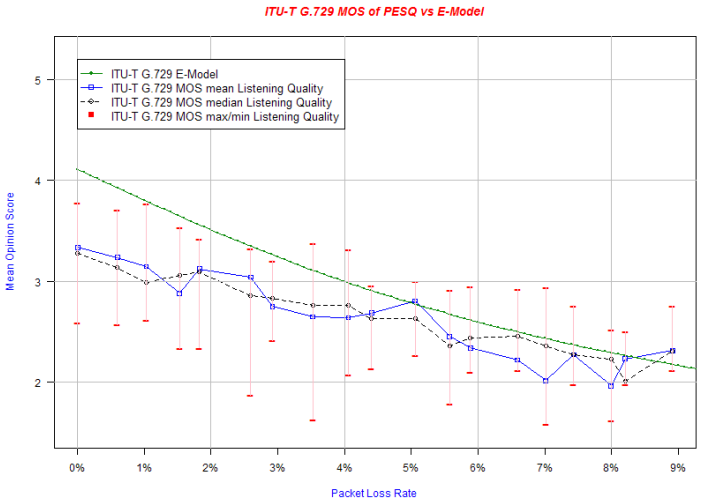


Figure 3.6: Average MOS of ITU-T G.729 PESQ and E-Model

average value of a group by eliminating those outliers to least at some extent. Moreover, as the data is right skewed, the median provides a better estimation.

Fig. 3.6 shows comparison results for MOS of PESQ and E-Model for ITU-T G.729 codec. The loss range goes beyond the theoretical maximum 8% loss rate because of the randomness of loss distribution that could lead to more packets lost in tests with loss rates $\geq 7\%$. The MOS results of PESQ vary largely near the E-Model score. This is because the intrusive assessment methods are heavily dependent on data difference between the original sample and degraded one. For speech samples with different languages, the content would make differences in final intrusive metrics. Packet loss is the key factor that may lead to varying results as a lost packet can happen in a speech period or a silence period. For these reasons, each vertical bar is a set of experiments that have similar loss rates and mean / median values are highlighted and lined in the graph. The results show correlation between PESQ and E-Model results despite of the content dependence of PESQ and randomness of packet loss generator.

Fig. 3.7 and 3.8 show the comparison results of PESQ and E-Model for ITU-T G.711μ and G.711a. Graphical representation holds the same explanation as for Fig. 3.6. One more point of analysis here is the cross comparison of different audio codecs using these results. While median values of real-time MOS for test-cases using G.729 are below the E-Model specification, MOS values rise above the E-model specifications

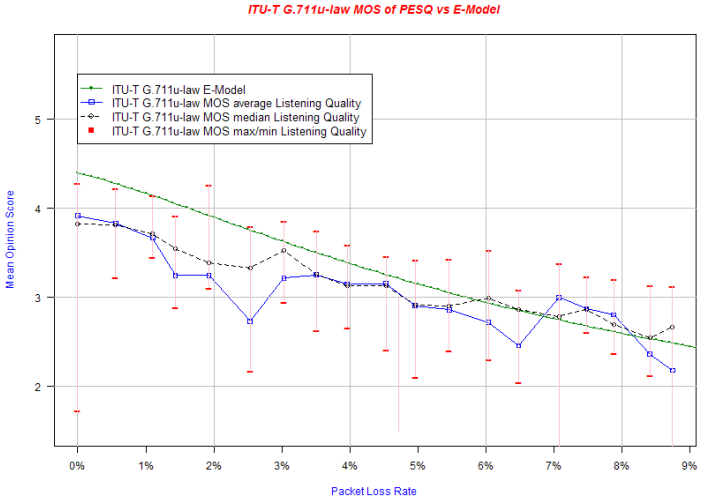


Figure 3.7: Average MOS of ITU-T G.711μ PESQ and E-Model

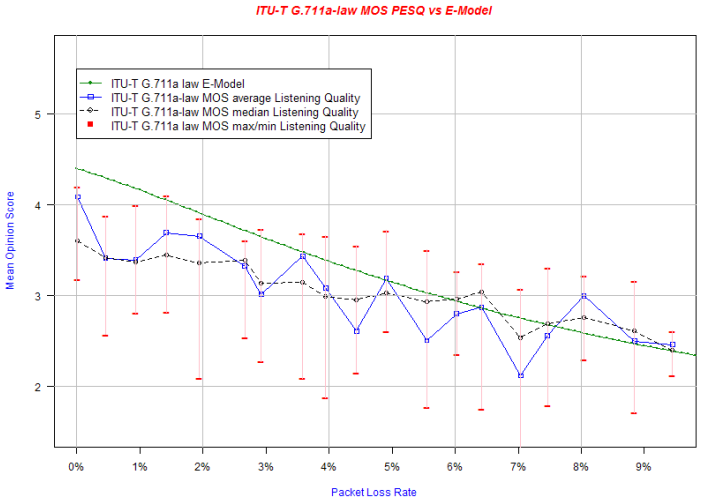


Figure 3.8: Average MOS of ITU-T G.711a PESQ and E-Model

for G.711μ and G.711a above 6% packet loss ratio. This observation infers that G.711μ and G.711a codecs perform better than G.729 for packet loss ratio greater than 6%. Thus, it gives an idea of selection of codecs depending on the network environment. Same experimentation could be done using various other codecs and a comparative analysis could be performed using the framework proposed to specify the best codec available for specific network conditions. This not only would assist the process of codec selection but also improve overall QoE, user experience and help in network planning for the product deployment.

3.6 Summary

We have presented a novel framework for testing audio and video phone services in enterprise networks. We have mentioned the need of a testing framework for enterprise because of the network and traffic specifications and requirements. Two different approaches for testing and their pros and cons have been discussed, one intrusive or off-line and one non-intrusive or in-line technique. We have presented the need of external tools and internal plugin for precise testing of VVoIP system's performance. Finally we have presented the test-bed set-up and few results to show the sample metrics and graphs used to evaluate the performance of AV telephony.

Chapter 4

Video Codec Performance

4.1 Introduction

Quality of service for video telephony could not be defined by network parameters only. For instance, there could be long delay, though the image quality is perfect and hence the subjectively video quality is excellent for the end-user or for some applications, resolution matters more than frame rate and vice versa. Thus, the quality of video as perceived by the end user is not dependent on the errors in image sequences but it varies from application to application. For the purpose of quality estimation of video phone services, perceptual evaluation of the video calls is necessary.

This chapter is dedicated for the discussion on different metrics used for quality assessment of video. There have been research, discussions and surveys on different types of metrics, their quality assessment algorithms and methodology, their merits and demerits for being used as quality assessment for different types of applications and their scope of being close to perceptual evaluation of video quality. Section 4.2 discusses various video quality evaluation and estimation metrics being used in this research. This section highlights the difference in methodology of quality assessment by these metrics and also the purpose of each metric to be used for specific applications.

The main objectives of this chapter are;

1. Implementation of a sample framework for quality estimation of video phone applications;

2. Performance evaluation of an enterprise communication software application in an enterprise network;

3. Defining procedure for analysis of results generated from the proposed framework in order to evaluate the performance of video phone application software;

4. Comparative analysis of different video phone applications using the proposed framework.

This chapter also presents results of quality assessment of video telephony using the framework discussed in Section 3.2. Section 4.3 describes a sample implementation of framework used to produce results presented in this chapter using to different video phone applications. Section 4.4 presents video quality evaluation results and analysis of those results. It also presents a comparative analysis of two different video phone applications to highlight the application of this work in real industrial scenario. A brief discussion on accuracy analysis of present standard, G.1070 Opinion Model for video telephony Rec (2007) is also presented based on the results presented.

Next section discuses various video quality evaluation and estimation metrics.

4.2 Video Quality Evaluation Metrics

There has been an evolution of quality metrics for video quality assessment Winkler & Mohandas (2008). PSNR and MSE have been the most popular metrics to evaluate the difference between two given images. While PSNR is another representation of MSE in logarithmic form, it computes quality by byte-by-byte comparison of the reference and processed images. PSNR does not address the properties of pixels, their correlation with each other and the relationship of different parts of image with each other and the interpretation of objects within the image by human vision. Instead, it simply computes the difference between images overlooking the type of content and data lying within. This content-agnostic property of PSNR makes it difficult for use as a perceptual metric. Two images might have same PSNR but could be far different from each other from a user's perspective.

To overcome this drawback of data insensitivity, feature oriented metrics were developed which took care of the spatial properties of objects lying within the image. These metrics take into consideration the impact of distortion as well as the content, on perceived quality. They are based on extraction of specific features and artefacts within the image. These features include structural information within the image like

edges, contours or object shapes like image blur. Artefacts could include special distortions produced by codecs or transmission link or some video processing steps like block errors. SSIM is one of the most popular metric in this category Wang *et al.* (2003a), Winkler & Mohandas (2008), Serral-Gracià *et al.* (2010). SSIM index computes the mean, variance and covariance of small patches inside a frame and combines the measurements into a distortion map. Motion estimation is used for weighing of the SSIM index of each frame in a video. It is regarded closer to human vision perception because of its resemblance to capture the human visual interpretation of structural boundaries and gradients.

Furthermore, research and subjective tests were done to develop a metric which emulates human perception of image viewing. Video quality metric (VQM) was developed as a part of this work Pinson & Wolf (2004). VQM divides sequences into spatio-temporal blocks, and a number of features measuring the amount and orientation of activity in each of these blocks are computed from the spatial luminance gradient. The features extracted from test and reference videos are then compared using a process similar to masking. VQM is used as a standard full reference metric by academic and industrial research community for quality evaluation of the video streaming systems.

ITU-T has been working on Opinion Model for multimedia quality assessment Rec (2007). This mathematical model as explained in section 2.3.5 is based on subjective tests and modelling human interpretation of impact of various factors on video quality. It captures the effect of network impairment on video quality in term of degradation caused by packet loss ratio. Moreover, it also considers the impact of codec implementation viz. frame-rate, bit-rate, codec robustness factor against packet loss and optimized bit-rate and frame-rate for a given bandwidth. However, it does not consider the importance of video content and packet loss patterns on perceived video quality. Current research discussed in Section 2.8 suggests modifications in Opinion Model for video telephony based on video content. These modifications are made only at 0% packet loss and hence do not fully account for packet loss scenarios.

Next section discusses the implementation of the framework discussed in Section 3.2 for video telephony. The suite of tools used for this implementation have been discussed in Section 3.3.

4.3 Testing Framework and Implementation

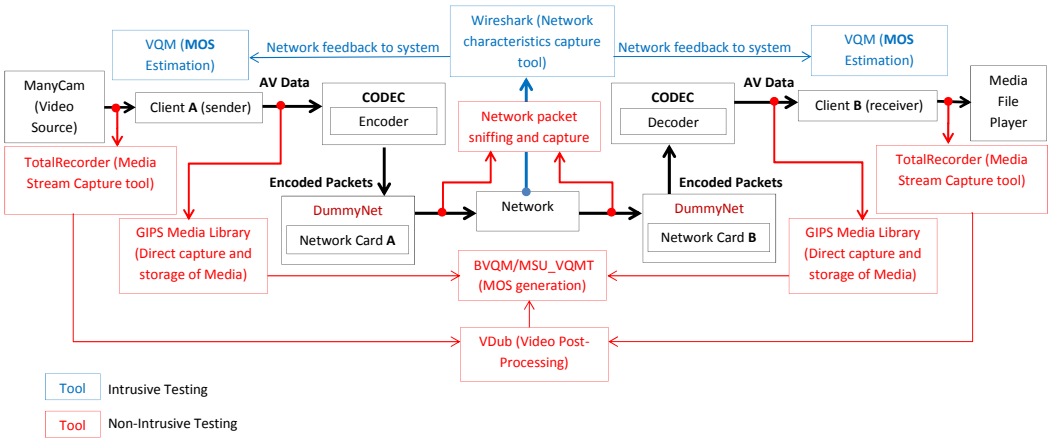


Figure 4.1: Implementation of the Quality Assessment Framework for Video Telephony with various tools and utilities used in its realisation, showing: the VVoIP system and under test (black), non-intrusive QoE assessment (blue) and intrusive QoE assessment (red).

Video QoE in particular, depends on the implementation of codec within the application, the network environment and the variation in network resources available. A sample implementation of framework specific to evaluation and analysis of video QoE is presented in Fig. 4.1 and has the following functionalities;

- It works as an external plugin to the application to run sets of tests under varied network and application environment conditions;
- It collects data at 3 different points in the data flow process and supports both intrusive and non-intrusive testing methodologies;
- It support various QoE metrics depending on type on analysis, suite of tools and resources available;
- It performs manual as well as automated testing under varied network conditions to test performance of codec or service.

An implementation of framework as shown in Fig. 4.1 is specific to video telephony clients and is used for testing and comparative analysis. Each end of the call has an application client with the framework plugin specifically configured for the application. Either end could be used as a sender or receiver. The experiments were conducted in real enterprise network provided by IBM with real video phone clients. Two different clients used here were *IBM Sametime 8.5.2 IFR* ibm (2013) and *Skype 6.6.0.106* sky (2013).

Fig. 4.1 represents only an implementation of the proposed framework to evaluate the quality for the two applications mentioned and should not be taken as a generalised implementation. The implementation and usage of tools could vary depending upon the application's input/output data format, decoder type and accessibility of the raw encoded stream at the receiver end. For example, in the case of hardware based decoders such as those present in some android devices, the tools for extraction of data and the data abstraction format could change and appropriate quality assessment metrics could be chosen but the basic methodology of testing and comparison of products would remain the same.

For two different clients we have used two different approaches, both of them involving intrusive testing and verification against non-intrusive results. One is automated using the plugin connected to the client and other one is by the manual process using a suite of tools. Fig. 4.1 representing the flow of data within the framework includes client specific components in black nodes, intrusive testing components in red nodes and non-intrusive testing components in blue nodes.

For sending a video file as a standard source so that the same experiment could be repeated with the same input, we use the video camera emulator software *'ManyCam'* as shown by node 'ManyCam' in Fig. 4.1 at the start of data flow process. *'ManyCam'* is selected as the input device instead of the integrated camera and hence our video source is same for each experiment and it appears as real-time interactive traffic is flowing from one end to the other. This component eliminates the device generated impairment in video source for each call. For the source we have used a standard .avi video file of 10 sec duration. The video test samples are provided by Recommendation ITU-T Rec (2008) as test signals for use in telephonometry. This recommendation describes that these test signals are applicable to various aspects of telephony products

and the signals are used as test samples for the experiments in order to eliminate the risk of choosing inappropriate samples that lead to inaccurate results.

TotalRecorder was used as the external tool to capture the display window of the received video and store it as desired video format and then send data for post-processing to other node. *VirtualDub* was used for post processing of video files such as frame synchronization and resolution settings. The advantage of this method is that it represents the data in the closest form of user interaction and captures the real image seen by the end user and hence results in overall perceptual evaluation of the system. In another approach, media data is captured within the client before encoding and after decoding as shown in Fig. 4.1 as node "Data capture of media stream from client". This provides us the actual codec impairment brought into the system if the network is supposed to be clean and does not require synchronization of any media property. This functionality was implemented using native functions of GIPS media framework available within the application client.

DummyNet Luigi (1999) was used to emulate the network and generate packet loss, where for non-congestion related drops, probabilistic match option is used to emulate links with uniform random loss patterns. This ensures that the packet loss is not content dependent and could happen at any frame or packet of media sequence providing randomness of network and relates to real-time traffic emulation. IP packets are captured from the network card of the end-user equipment as shown in Fig. 3.2, node "Network Characteristics Capture Tool". This helps in computing actual packet loss ratio, delay and jitter realized during each call. These statistics are fed into the standard mathematical models to estimate MOS using non-intrusive methodology as shown in Fig. 3.2 node "VQM (MOS Estimation)".

For our quality assessment and analysis, we have chosen PSNR and SSIM as the objective metrics for analysing video streams in the intrusive test methodology and Video Quality Metric (VQM) from opinion model for video telephony as objective MOS for the non-intrusive test methodology. As PSNR and SSIM are content dependent Joskowicz *et al.* (2011) and VQM is network and codec dependent Rec (2007), the set of parameters used for comparison are: sample contents (degree of temporal and spatial variations in video sequences) and packet loss rate. Two different video sequences were chosen as video sources having different spatial and temporal properties and thus having different contents. Thus, contents and network variance can be tested under the

designed experiment that keeps one variable fixed and changes the other one. *BVQM* McFarland *et al.* (2007) was used to compute the PSNR and *MSU-VQM* Vatolin *et al.* (2009) was used to compute SSIM of the degraded video streams using full reference method for comparison. Later these PSNR scores were converted to MOS in the scale of 1-5 using a linear mapping defined in the next section.

Comparative analysis of different systems or products is done on the basis of their relative changes in quality metrics with respect to changes in the packet loss rates. Different metrics represent different features and their behaviour with respect to external parameters changes. Our comparative analysis incorporates three different metrics to compare applications with different aspects and then presents a verified, composite analysis of the performance of the applications.

For results considering PSNR, the most important factors are the highest and lowest PSNR values recorded by the applications under the same packet loss range. Moreover, the change in PSNR with increase in PLR also represents the implementation or robustness of codec implemented. Without loss of generality, this work infers that the system with lower gradient loss in PSNR values and with higher highest and lowest values of PSNR shows better perceived quality as compared to the other.

SSIM index on a broad level, represents the distortion values with details of structures and their movement inside the received video sequence. The higher value of SSIM for a given PLR shows less distortion and higher precision of structural display within the video sequence which is closer to human eye perception. This work infers that higher the SSIM index, better is the QoE response of the corresponding application.

Video MOS is the objective QoE metric to measure the perceptual quality of received video. Here MOS is derived from PSNR scores using equation 4.1 based on table 4.1 presented in next section. The most important aspect of MOS is the range of change of the MOS and the slope of MOS curve with respect to the PLR. Narrower the range of MOS change, more robust is the codec to network loss. Moreover, higher the MOS value, better is the user experience with the received video. Change of MOS with increasing PLR shows the actual response of the application to network losses. This work infers that higher the MOS, less range of change of MOS, less slope of MOS curve with respect to PLR in the region of interest, shows better performance of the application with respect to the user perceived quality.

4.4 Results and Analysis

The experimental results are the PSNR, SSIM and MOS values of the point-to-point video call test-cases under varying packet loss and source contents using two different video phone applications. The tool used to generate the packet loss results in a normal distribution, centred at the requested loss rate value. And thus even if the specified value of packet loss viz. 5% for a video call containing 600 packets should result in expected number of packets lost to be 30, it could vary from 25 to 35 resulting in packet loss to be 4.2%-5.8%. In some extreme cases, the deviation may be larger. Thus the theoretical value of packet loss set in the tool is not used while computing call quality rather the real packet loss rate is calculated by analysing the video RTP stream using the packet capture and analysis tool.

Two different video phone applications were used to get video quality results. First one is an enterprise unified telephony solution *IBM Sametime* ibm (2013) and the second one is a popular free audio-video phone application *Skype* sky (2013). Performance results have been presented in terms of PSNR, SSIM and MOS. A comparative analysis of performance of these two application is also presented. This depicts the applicability of proposed framework onto commercially available video phone solutions.

For testing *IBM Sametime*, each set of experiment included 17 calls with varying packet loss. 15 sets of tests were performed and in each set, packet loss rate was set from 0% to 8% in 0.5% increments. For video sample, video media sequence was of duration 10 seconds, 30 fps frame rate, 384Kbps network bitrate, QVGA resolution and the codec under test was H.264 AVC. While testing *Skype*, the same set of experiments were involved with 5 sets of experiments performed. In this case, the codec under test was VP8 and all tools were used manually.

Next subsection presents quality results in terms of PSNR and a comparative analysis of presented results including both applications.

4.4.1 PSNR Results

Fig. 4.2 and Fig. 4.3 represent the results of quality testing of *IBM Sametime* and *Skype* in terms of PSNR with same video source and under the environment conditions as specified above. For the ease of representation of data, we have divided packet loss ratio axis into bins of width 0.5 unit, centred at increments of 0.5 for *IBM Sametime* and

width of 1 unit centred at increments of 1 for *Skype*. PSNR values represented here are mean and median values of the points lying within these bins. Also the regression plot of the whole data set is plotted using least mean squared difference. These graphs give us an idea about the expected behaviour of video phone application under the packet loss scenarios and with source content having similar range of spatial and temporal activity within them.

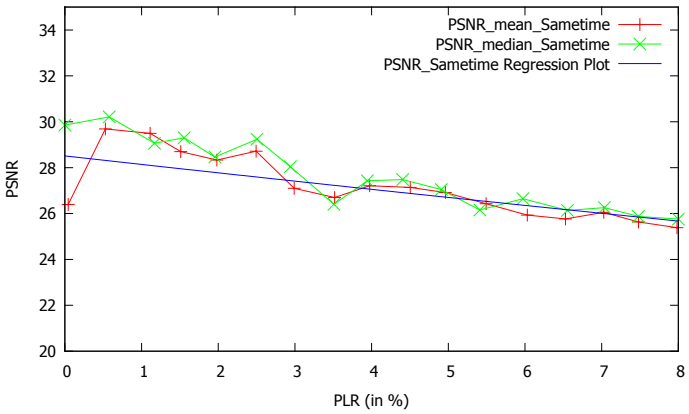


Figure 4.2: Average PSNR of *IBM Sametime* versus PLR

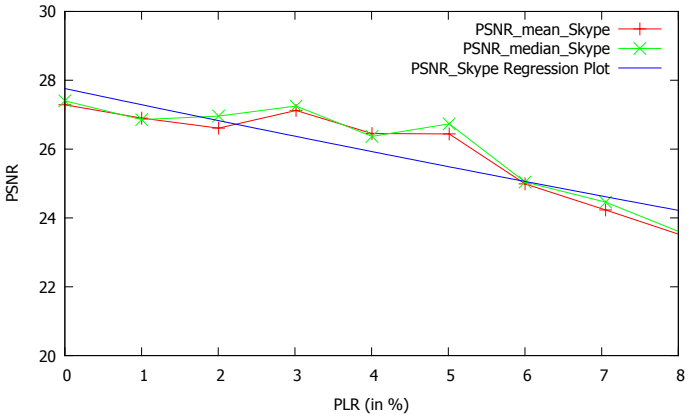


Figure 4.3: Average PSNR of *Skype* versus PLR

Fig. 4.2 and 4.3 show the performance of *IBM Sametime* and *Skype* in IBM Lab network environment having bandwidth which is more than sufficient for video stream transmission over any network. PSNR values were calculated for each call's video sequence and we observed that the data was right skewed. Exponential regression model was adopted to plot a fitted curve for the scatter points against packet loss

ratio in percentage. The packet loss was calculated by using actual number of packets received and lost during the call. The graph follows exponential decline to left as expected, by increasing packet loss, the video quality decreased. The value of PSNR at 0% loss is not above 45 dB because of the coding distortions and degradation in data due to interaction between different media frameworks in the source code of client application.

We infer that not only network and codec distortions affect the overall media quality but the additional layers of abstraction in client code and their interaction with data also plays its role in defining the end point QoE for video phone applications. Different types of packets contain different types and amount of data used for the reconstruction or decoding of the video's image frames and hence under random packet loss there is a variation in PSNR values at a given packet loss ratio. If some important packets are lost which contain reference to other packet's data as well, then PSNR drops drastically. We infer that video quality depends heavily on the packet loss patterns as well and consider this issue in Chapter 5.

Using these graphs, performance of both systems could be analysed. While median values of PSNR for *IBM Sametime* vary between 31dB to 25dB, they acquire slightly lower values for *Skype* which is between 27dB to 23dB. The degradation of PSNR from higher to lower value with increase in packet loss ratio also conveys the robustness of codec implementation within the application against packet loss. While in Fig. 4.2 data points are right skewed, with nearly exponential decline, Fig. 4.3 shows that the system is countering packet loss at PLR lower than 5%. Fig. 4.3 shows that system tries to maintain nearly constant video quality as shown by the low range of PSNR change (26.5dB to 27.5dB) for PLR below 5%, but PSNR values drop drastically as PLR increases more than 5%. This shows that this application is designed to take care of lower packet loss probably with some error concealment algorithm but not for higher packet loss in critical network conditions. We affirm the behaviour shown in graphs and their analysis, with our subjective test experience with the given applications. However, *IBM Sametime* does not use error-concealment algorithm.

A sample of the original and degraded frames are produced for both applications at a given packet loss ratio. Fig. 4.4 and Fig. 4.5 show a sample image frame sent form *IBM Sametime* and *Skype* respectively and their degraded copy received at the other end under 5% packet loss.

Figure 4.4: Original and degraded video frames transmitted at 5% packet loss form application using *IBM Sametime* resulting in PSNR=26.22

Figure 4.5: Original and degraded video frames transmitted at 5% packet loss form application using *Skype* resulting in PSNR=24.49

Fig. 4.4 and 4.5 clearly show the difference in video quality for the two given applications. Same difference is shown by the corresponding PSNR values of these clips.

4.4.2 SSIM Results

As our effort to bring in perceptual evaluation of video phone services, we have also considered SSIM as one of the important metrics for video quality assessment as it is closer to human eye perception in comparison to PSNR. SSIM values for the dataset collected from the above mentioned applications are shown in Fig. 4.6. It shows the SSIM values for every call under test and also presents a regression plot of the dataset based on least mean squared difference algorithm. As clear from the figure, it shows

high correlation with our framework results. Data is right skewed with random variation against packet loss ratio. SSIM index is higher for *IBM Sametime* as compared to *Skype*. This observation is in correlation with our results in terms of PSNR. This verifies the working of our framework but the degree of preciseness could not be measured without comparing against a set of subjectively tested video database.

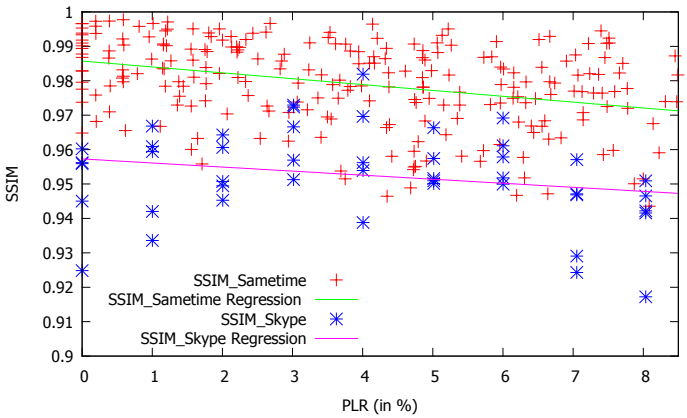


Figure 4.6: SSIM Regression plot of ITU-T *IBM Sametime* and *Skype* versus PLR

4.4.3 MOS Results

Fig. 4.7 shows the comparison of the MOS derived from PSNR using the algorithm in table 4.1 derived by Ohm (2004) and opinion model for video for applications using H.264 and VP8. We have divided packet loss ratio axis into bins of width 0.5% and centred at increments of 0.5 viz. 0%, 0.5%, 1% and so on. Within each bin, there are several data points and each data point presented here in the graph represents the mean and median of those points inside the bin. In Equation 4.1, we interpolate between the values presented in Table 4.1 by assuming that the relation between MOS and PSNR inside these regions is linear.The median values gives us a better, smooth and relevant graph because the data is right skewed and median is more resistant to skew. Hence, we have shown comparison of the MOS generated by opinion model at given packet loss rates and the median values of MOS derived from PSNR within each bin. We can see that there is a difference in the estimated and observed values of MOS. We infer 2 major reasons for that. First one is that, the opinion model is not accurate enough to predict MOS, given a codec implementation at a given frame rate, bit rate, source content and

PSNR(dB)	MOS
>37	5(Excellent)
31-37	4(Good)
25-31	3(Fair)
20-25	2(Poor)
<20	1(Bad)

Table 4.1: Conversion table from PSNR to MOS

packet loss ratio. The other one is that, the system is not actually performing to its expectation and there are some bugs causing the degradation to media or there is a need of some advanced techniques to be deployed to benchmark system's performance to the expected one.

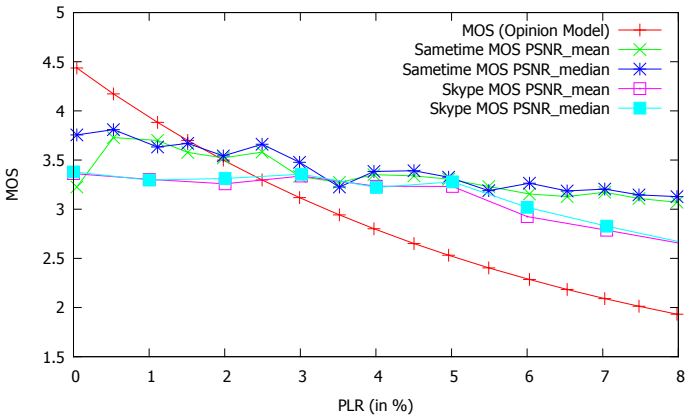


Figure 4.7: Comparison of average MOS of *IBM Sametime* and *Skype* and Opinion Model

While the same analysis persists for comparing MOS of two different applications as for PSNR, we can verify the steep decline of call quality for application using VP8 above 5% packet loss rate from Fig. 4.7. We also observe a difference between MOS values as derived from PSNR and from opinion model. This observation is critical for accuracy analysis of the opinion model of video telephony. While there is no generic model for all codecs, there are no standard coefficients present for different codecs as well. The coefficients used here are specific to H.264 as derived by Joskowicz *et al.* (2011) and thus are not specific to VP8 codec for MOS verification.

$$MOS = \begin{cases} 5, PSNR > 37 \\ 0.15 * PSNR - 0.65, 31 < PSNR < 37 \\ 0.153 * PSNR - 0.813, 25 < PSNR < 31 \\ 0.184 * PSNR - 1.673, 20 < PSNR < 25 \\ 1, PSNR < 20 \end{cases} \tag{4.1}$$

4.5 Summary

Here we have presented a testing framework for assessment and analysis of performance of video telephony application software in enterprise against fluctuating environmental conditions. We have discussed the need of a perceptual quality assessment of video telephony products owing to their dependence on network fluctuations and codec implementations. Moreover, due to the resource requirements and network and traffic specifications of the application, pre-analysis of the product before its deployment is essential for optimized resource allocation during its run-time. We have discussed two different testing methodologies depending on the nature of testing and input parameters for quality assessment—intrusive and non-intrusive. A sample implementation of our framework using the given test methodology is also presented, including a suite of tools used for specific purposes.

Finally we have presented the results and analysis of quality testing of two sample video telephony applications under industrial lab conditions. One being *IBM Sametime* and the other *Skype*. The results include visual graphs including widely used industry standard quality evaluation metrics against packet loss scenarios. Analysis of products depending on different metrics is presented and the need of enhancement of the current testing procedure is also highlighted. Our results with two different methodologies correlated well, however in some cases where a variation exists, highlight the need for adjustments to the standard model. High correlation between the testing methodologies verify the performance of our framework.

Our framework is limited to black box testing of video phone clients, which is somewhat limited. The need of extension to our framework to consider the effects of distortion in real-time data by abstract layers of coding within the videophone client, has come up as conclusion of our present work and is planned to be considered into our future work. Moreover, future work consists extension of the present framework for multi-party and web-conferencing systems and including more quality degradation

parameters like video content and packet loss patterns. Preciseness of the opinion model needs to be tuned depending on the resolution of video, video content and type of packets lost during call. Specific frame types and their impact on the overall QoE is also an important point to be considered in our further research.

Chapter 5

Dependance of QoE on Packet Loss Pattern

5.1 Image Slicing in Video Sequences

Previous chapters have shown the actual limitations and restrictions in measuring the QoE of video phone services as well as the effects of network impairments on the overall call quality due to degraded video sequences. Delay, bandwidth and packet loss ratio are the major parameters to affect the video quality during transmission through a network. Albeit network impairments are key dependencies of video quality, they are not the only one. Quality of video transmitted heavily depends on the content of video as well the type of packets lost. Typical uncompressed video bit rates are 270 Mbps for standard-definition (SD) and 1.485 Gbps for high-definition (HD) Greengrass *et al.* (2009a). Network bandwidth constraints make streaming such high-band video contents from end to end impractical. So we use compression in the form of video encoding.

As discussed in Chapter 2, compression techniques divide the video sequence into individual images, then subdivided parts of the image are called slices. Slices are further divided into macro-blocks and blocks which adds one more level of complexity to the system. The loss of a specific type of frame or slice or macro-block may produce different types of distortions into the video sequence and that too at some specific places in the individual image as well as in the whole video sequence. The encapsulation of these differently encoded picture frames into fixed length IP packets also produces another

issue that losing one IP packet may cause loss of data for a single frame or for multiple frames. Therefore, it is important to address the effects of packet loss patterns on video QoE or effects of loss of specific kind of image frames on video QoE.

Motivation of this work lies in the fact that every packet is not the same. Every packet holds its own importance in reconstruction of the whole video sequence by the decoder at the receiver. Interdependence of packets on each other to decode the video is also an important aspect of decoding. These aspects give priority to some specific packets over the other in terms of their impact on overall video QoE. Hence we need to identify the quality variations depending on the packet loss patterns and which packets should be considered more important than the others to optimize the video QoE for the end-user.

The main objectives of this chapter are;

1. Introduction of methodologies to capture type of frame/slice lost during a call and classification of test-calls based on the type of frame/slice lost for the purpose of comparative analysis of MOS based on the type of frame/slice lost.

2. Study of impact of specific frame loss on the overall quality of the video phone application.

3. Study of various factors apart from network impairments which could impact the quality of the video phone and thus the overall quality of the communications software.

Different types of frames contain different data and dependencies on each other and have different level of information content. This information content is one of the key aspects of this research, which defines the priority of frame type, depending on the information contained in it and its impact on the overall QoE of the video service. One frame lost may produce different artefacts as compared to some other one. This chapter defines a test-bed to run experiments on an AV client in a real industrial environment under varied network conditions. These experiments give an overview of the impact of different types of frames on the QoE. Following sections include a study on behaviour of random frame loss and produce comparative graphs showing QoE under varied network loss as impacted by typical frame loss.

In this chapter, Section 5.2 includes the features of the test-bed used as an extension to the proposed framework in Chapter 3 for assessing the impact of specific packet loss on overall QoE of video streaming services and in particular a video phone service. Section 5.3 explains the two different data classification methodologies used for categorising test-calls into different classes for the purpose of comparative analysis depending on the type of slice(s) lost within the call. Section 5.4 presents the results and comparative analysis of the different classes. Moreover, how the packet loss patterns or loss of a specific type of frame affects the overall video QoE is studied and discussion is presented. Section 5.5 summarises the work in the chapter.

5.2 Experimental Testbed

As an objective to determine the impact of specific frame loss on overall QoE of video phone service, this section targets at developing an experimental test-bed based on the testing framework described in Chapter 3 with some extension for some specific purposes concerning the experiment. The principal objectives and features of the test-bed apart from the previously specified features of the proposed framework are;

- It is capable of identifying the exact number of packets lost during transmission and the type of packets lost containing specific information about the video transferred;
- It supports extraction of data in terms of the number of specific slices lost within each call and categorizes each call into defined classes for the purpose of comparing MOS of different classes defined.

A sample implementation of the used test-bed is shown in Fig. 5.1. As compared to the previous test-bed described in Chapter 3 and 4, this one focusses on collecting network data and identifying the packets lost. Hence UDP traces have been captured on both sides of the call to compare the packet sequence numbers of the packets lost and gather the information lost while transmission. Test calls are run as explained in previous experiments. Variation in network statistics is automated within the framework and could be manipulated to vary on per call basis to generate random packet loss patterns. The test-bed consists of a set of tools and video telephony client on the two end-points, each having the embedded testing framework as a plugin.

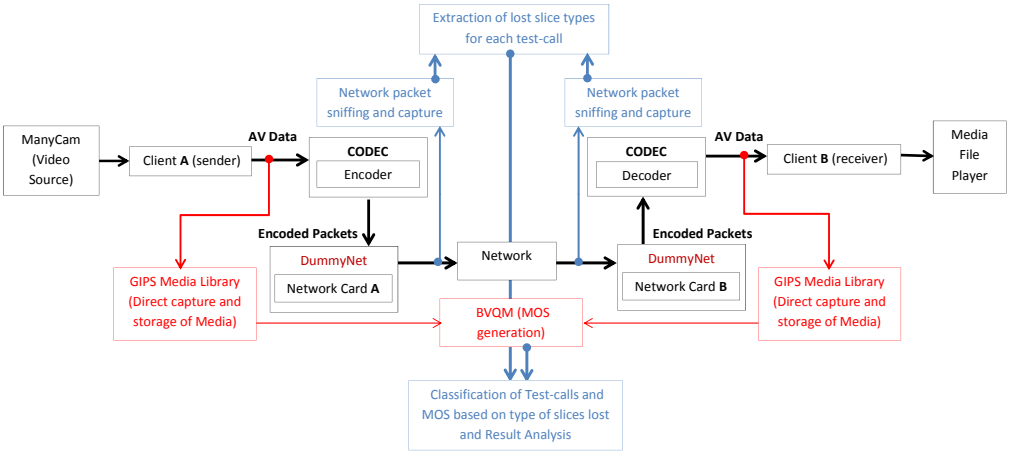


Figure 5.1: Implementation of experimentation test-bed, representing flow of data, tools used, showing: Quality testing framework from Chapter 3 (black and red) and extension to the frame for the purpose of assessing impact of slice loss on video QoE (blue).

Virtual camera and a standard video source have been used as explained in Chapter 4. DummyNet Luigi (1999) was used to emulate network and generate packet loss, where for non-congestion related drops, probabilistic match option is used to emulate links with uniform random loss patterns. This ensures that packet loss is not content dependent and could happen at any frame or packet of media sequence. This ensures the randomness of the network and relates to real-time traffic emulation.

GIPS media framework was used to capture the video sequences within the video phone client. This method eliminates the synchronization problem as the capturing of data on both ends is synchronized with data control over the transmission link. Video sequences were stored in .avi format.

CVQM was used for quality metric generation with full-reference calibration. PSNR was generated as the quality metric and then later converted into MOS using equation 4.1.

Extension to the proposed framework was in the form of additional computational node involving Wireshark. This node carried out the tasks to capture UDP packet traces and then decoding them to H.264 packets. Lost or dropped packet sequence numbers were filtered out from the receiver side and then compared from the traces on sender side to identify the amount and type of the data packets lost. Wireshark

provides feature to lookup for specific **slice_type** in all packets and thus it was used to determine packets containing specific type of slices lost within a call.

The experimental results are the MOS values of the point-to-point video calls under varying packet loss and source content. The tool used to generate the packet loss results in a normal distribution, centred at the requested loss rate value. Thus the theoretical value of the packet loss set in the tool is not used for computing call quality, rather the real packet loss rate is calculated by analysing the video RTP stream.

5.3 Data Classification Methodologies

For video, 15 sets of test-cases were performed and in each set, PSNR was computed against packet loss rate ranging from 0% to 8% at 0.5% increments. The video media sequence was of 10 seconds duration, 15 fps frame rate, 384Kbps bitrate, and QVGA resolution and the application under test was *IBM Sametime* using H.264 AVC as video codec.

Two different methodologies were adapted for categorising the test-cases into different classes depending on frame type lost and to assess the impact of frame loss of the video QoE. The first one involved classification of test-cases based on the dominant impact of a particular type of slice loss. This method used manual classification of each call. Another method involved K-means sampling Hartigan & Wong (1979) of the test-cases based on the type of slices lost and clustering test-cases into the nearest class or slice type. Both the methods and their corresponding results are presented below.

5.3.1 Manual Classification

As the initial step of classification, the input parameters needed are sequence number of IP packets lost, type of slices lost within those IP packets and the amount of each slice-type lost within each test-call. These statistics were collected using a comparative analysis of sent and received packets traces using Wireshark. Lost packets on the receiver side were identified by the breaks in series of packet sequence numbers. These lost sequence numbers were filtered out of the sent packet trace and this filtered data represented the data actually lost while data transmission. Contents of the refined IP packets were monitored using Tshark and their NAL headers were segregated. NAL headers were used to extract information about the type of payload inside the NAL

unit. Here again, packets containing coded data slices were filtered. After this step the filtered data set contained the NAL units of the coded data slices, lost during the transmission. Information about type of slice could be extracted by looking into the **slice_type** parameter using Tshark.

A sample representation of the required dataset is shown in Table 5.1, after the initial extraction of type of slices lost for each test-call. For the manual method of classification of test-cases, 6 different classes were defined; I, P, B, IP, PB and IPB and the test-cases were classified depending on the type and amount of slices lost. For instance, a test-case which has lost only I slices, would be categorised under I class and similar for P and B-class. Test cases with multiple slice type loss viz. I and P would be categorised under IP-class, and same holds for PB and IPB-class. However test-cases with loss of multiple slices and having one slice loss as dominant would be classified as single slice loss. For instance, a test-case which has lost 3 I-slices and 1 P-slice was categorised under I-class since the impact of losing 3 I-slices is theoretically far more than losing 1 P-slice. This process gives power of discretion into the user's hands and thus can not be considered reproducible. However, due care was taken during categorising the data in all fairness and generality.

Advantage of this approach is that every test-case is subjectively observed and analysed before categorising it into a particular class. This method brings in the subjective approach of impact analysis, which is closer to the perceptual evaluation of video quality. However, manual method involving human interception also exposes analysis procedure to human error and individual perception. Moreover, it does not consider the weights of the slice type lost in terms of the amount of individual slice type lost in case of multiple slice loss.

After classification, each class's data was separated, analysed individually and relevant graphs were plotted to analyse the impact of each class on overall QoE MOS. Relevant graphs and their analysis are presented in Section 5.4.

Fig. 5.3, 5.4, 5.5, 5.6, 5.7 and 5.8 plotted in Section 5.4 show the relevant data plots of the classes categorised using the manual process.

Manual classification method generates good results with each class defined to contain points belonging to the specific cluster. However, manual classification involves the risk of human error creeping into decision making. Moreover, with increase in size of data set, it is cumbersome to classify each point into different classes manually.

Packet Loss Ratio	I-Frame lost	P-Frame lost	B-Frame lost	SI-Frame lost	SP-Frame lost	Class Decision	MOS
0.386	1	0	0	0	0	I	3.2
1.957	0	3	0	0	0	P	3.5
2.204	0	1	1	0	0	PB	4.0
7.063	2	3	5	1	2	IPB	2.9

Table 5.1: Sample representation of final extracted data set from recorded data, showing packet loss ratio, types of frames lost and their numbers, manual classification of test-cases into I, P, B, IP, PB and IPB classes and MOS for each test-case.

The requirement of an automated algorithm leads to another method of classification which performs the same actions on data set to classify them in different classes and which produces results correlating with the manual classification. Basic fundamentals of manual classification are, checking the nearest cluster of the data point and then assigning it to that cluster. In mathematical form, this corresponds to clustering of data points based on the distance from each cluster. *Euclidean* distance classification involves nearly the same strategy as manual classification to categorize data points and thus is taken as a potential solution with promising future gains for automation of the process.

5.3.2 Euclidean Distance Classification

Second method of classification, which uses mathematical analysis of the extracted data, is used for autonomic classification of data points into classes. This method reduces the human interception in decision making, considers the weights of each slice according to their priority and importance and classifies data set into less number of classes: I, P and B only. This methodology is inspired from K-means clustering method Hartigan & Wong (1979), which defines partitions in a set of n data points, with each data point being assigned to a cluster with nearest mean. K-Means clustering defines means of clusters within the data points, however this analysis has not defined the means and cluster centres within the data set. This analysis assumes fixed axis as mean and drags the local points to the nearest axis for categorization.The three classes defined are I, P and B which also serve as the 3 axes of the 3D data set plot containing data points

with coordinates (i, p, b). Where i, p and b represent weighted sum of I, P and B-slices lost in the test-case respectively.

During manual classification of data points, due care was taken for importance of specific frame type viz. I frame being more important than P and B frame depending on the amount of data present and being the reference frame. However, manual process does not guarantee exact weights to be considered for each slice type. In the mathematical analysis, each data point in form of (i, p, b) consisted weighted sum of lost slices. Weights were calculated depending on the GOP structure. For the present case under study GOP structure used was 15:2 i.e. 15 slices in between consecutive I-slices and 2 B-slice in between every I or P-slice. Thus, assigning weight to each slice type on the basis of its frequency in a GOP gives I-slice 46%, P-slice 36% and B-slice 18% weight. Hence, for each test case, number of I-slices lost were multiplied with 0.46, number of P-slice with 0.36 and B-slice with 0.18 to form a generalised point in 3D space as (i, p, b). An example of such a data point is present below;

For a test-case with PLR 5.6 losing 2 I-slices, 1 P-slice and 5 B-slices, the final data point in the 3D space would be;

$i = 0.46 * 2,\ p = 0.36 * 1,\ b = 0.18 * 5$

$data\ point$ = (0.92, 0.36, 0.90)

Each test-case was assigned a data point in the 3D space according to the weights and the method presented above. The next step was to categorise them into different classes. For that purpose *Euclidean* distance of each point from each axis was used as the parameter to decide the class type. Axes of 3D plot were the three defined classes, I, P and B. Each data point was assigned to the class where the distance of axis was least. *Euclidean* distance is based on *Pythagoras* formula. According to *Pythagoras*, in a right triangle length of hypotenuse is equal to square root of summation of squares of other two sides. This work uses the same simple formula to calculate the distance of a point form an axis. For example distance of point (i, p, b) from X-axis would be;

$$X - distance = \sqrt{p^2 + b^2}; \tag{5.1}$$

i.e. by nullifying the x-coordinate of the point and considering a right triangle in y-z plane and calculating the length of hypotenuse using *Pythagoras* formula. Similar

PLR	I-Frame Lost	P-Frame Lost	B-Frame Lost	Decision	MOS
1.7	1	1	0	I	3.1
3.4	0	1	0	P	3.5
3.5	2	2	0	I	2.9
1.1	0	0	1	B	4.2
4.5	1	3	2	P	2.7

Table 5.2: Sample representation of test cases with classification done by mathematical analysis using weighted slice loss in 3D space having axes I, P and B.

operations were done on all data points and every point was assigned to a class represented by the axes I, P and B. Table 5.2 shows a sample representation and decision of classification of test-cases based on their weights of each slice type lost.

Fig. 5.2 shows a sample representation of classification of data points based on their *Euclidean* distance from different axes.

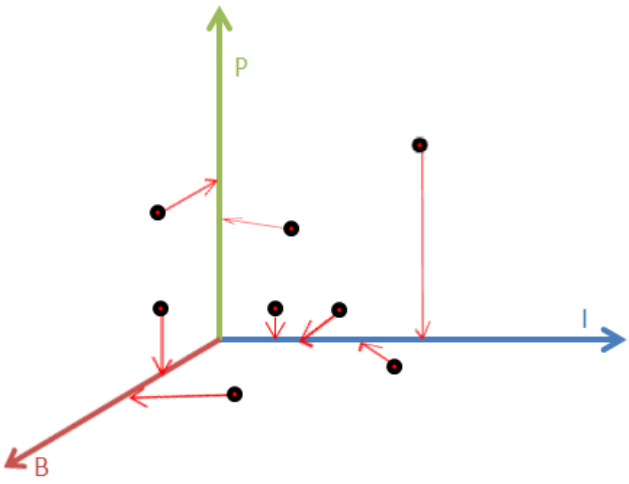


Figure 5.2: A sample representation of *Euclidean* distance method of classification of data points. The axes represent I, P and B classes. Black dots represent data points scattered in 3D space and red arrows represent their chosen classes based on their *Euclidean* distance from different axes.

Test-cases of each class were plotted separately to assess the impact of slice loss patterns on the overall QoE of video telephony. Fig. 5.9, 5.10 and 5.11 plotted in Section 5.4 show the MOS versus PLR graphs plotted for test-cases and their regression plots for I, P and B classes respectively considering exponential decline of MOS with increasing PLR.

Next section presents a detailed analysis of these results and the comparative study of the impact of specific slice type loss on overall QoE of video telephony.

5.4 Analysis of Results and Comparison

Previous sections have covered the test-bed set-up to assess the impact of specific slice loss on the overall QoE of video telephony and the classification methodologies for comparative analysis. This section covers the analysis of results presented and comparison of impact of different loss types on the overall video QoE during the experiment.

Two different methodologies for data extraction and analysis presented above, show high correlation with each other in terms of quality evaluation results. A detailed discussion on each graph is presented below.

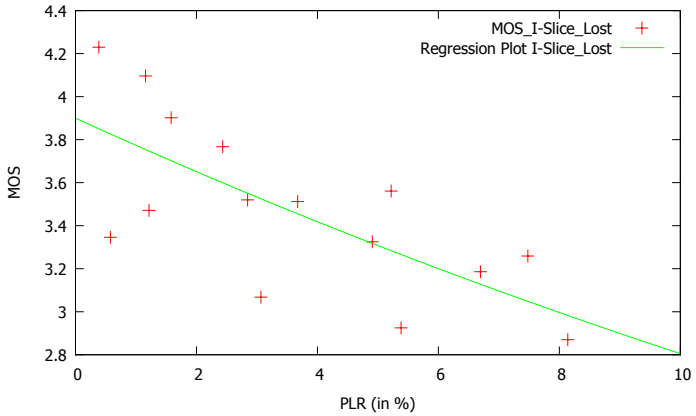


Figure 5.3: Test-cases categorised using Manual method with loss of I-slices only and corresponding PLR vs MOS plot and regression plot considering exponential decline of MOS

Fig. 5.3 shows the MOS versus PLR graph of test-cases which lost only I-slices during transmission. Out of 255 test-cases, only few of them have lost only I-slices and resultant regression plot of the data points shows steep decline in call quality. This

decline is in coordination with the theoretical explanation of functionality and features of I-slices. One I-slice lost does not only affect the present image but also takes away the reference point for the following P and B-slices. At the time of high temporal content in the video, losing I-slices have caused severe degradation into the whole GOP and the video scene occurs to be damaged. This observation has been verified in subjective testing of the same video sequences. The decline in call quality varies from an average of 3.9 to 2.8 which is difference of 1.1 unit MOS over a PLR range of 0%-10%. At high PLR, more than one I-slices have been lost and thus the MOS remains low. This graph infers that losing I-slice causes severe degradations and hence results in low MOS and skewed regression curve. This graph explains the importance of I-slices in determining the overall video quality.

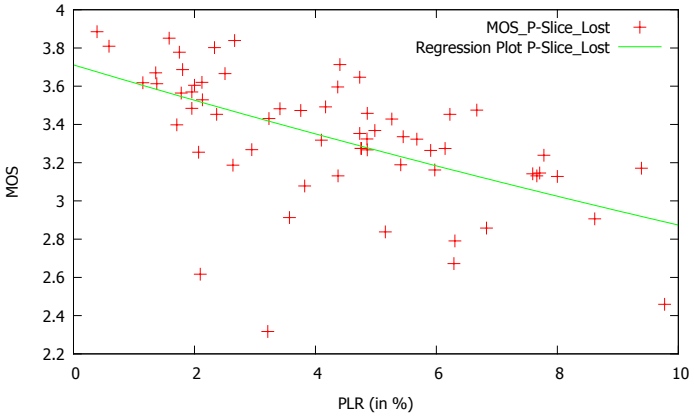


Figure 5.4: Test-cases categorised using Manual method with loss of P-slices only and corresponding PLR vs MOS plot and regression plot considering exponential decline of MOS

Fig. 5.4 shows MOS versus PLR graph for test-cases losing only P-slices. P-slices act as reference points for B-slices and other P-slices as well. Hence, losing a P-slice could impact the video sequence in small chunks of degraded video scenes. Here the gradient of loss for MOS is from 3.7 to 2.9. Although the range of MOS variation is less, the highest MOS is also less than the highest MOS for I-slices loss. Careful observation of the graph explains that most of the test-cases in this class start from PLR 1% which means that there has been a loss of more than one slice. Thus the amount of slice loss in this class explains the low estimated MOS at 0% packet loss. This observation has been verified with the data set present and also by subjective testing of the filtered

P-class test-cases.

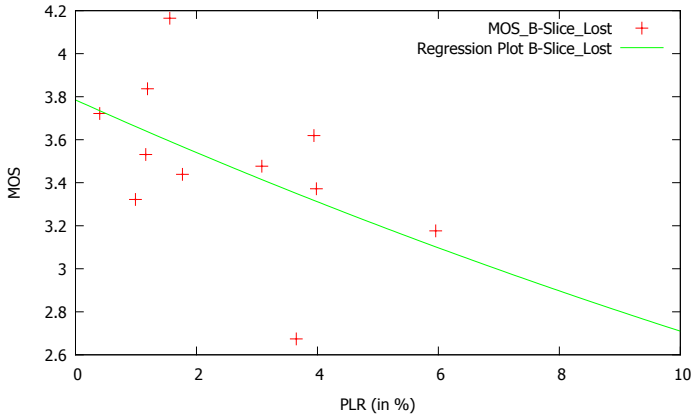


Figure 5.5: Test-cases categorised using Manual method with loss of B-slices only and corresponding PLR vs MOS plot and regression plot considering exponential decline of MOS

Fig. 5.5 shows the MOS versus PLR graph for the test-cases losing only B-slices. While the number of test-cases losing only B-slices are quite less as compared to P-class and I-class, the regression plot holds only approximation of estimated MOS at boundary values. The gradient of MOS decline against the increasing PLR is nearly same as P-class. Although here as well the amount of slices lost have not been shown in the graph, however more the slice loss less is the MOS.

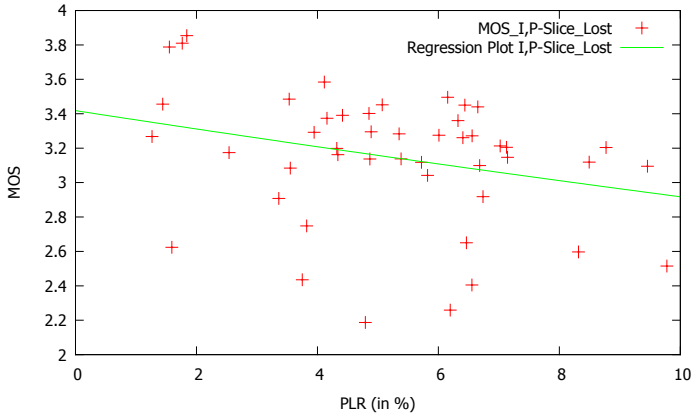


Figure 5.6: Test-cases categorised using Manual method with loss of I and P-slices only and corresponding PLR vs MOS plot and regression plot considering exponential decline of MOS

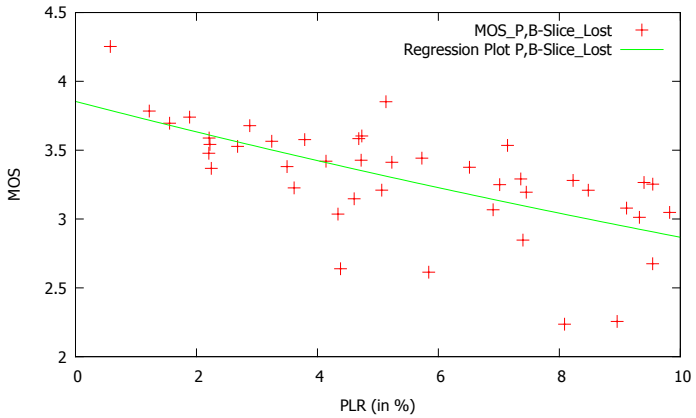


Figure 5.7: Test-cases categorised using Manual method with loss of P and B-slices only and corresponding PLR vs MOS plot and regression plot considering exponential decline of MOS

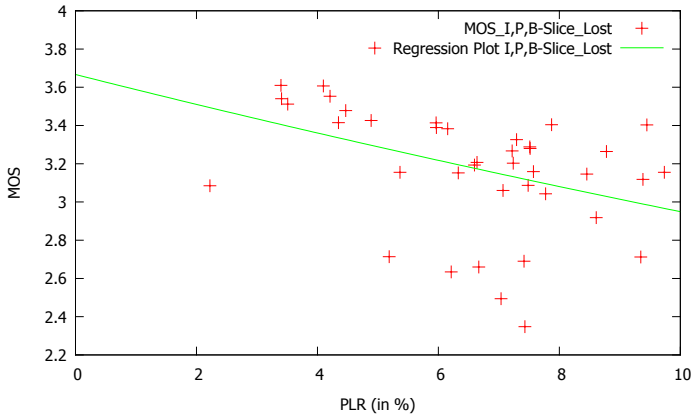


Figure 5.8: Test-cases categorised using Manual method with loss of I, P and B-slices only and corresponding PLR vs MOS plot and regression plot considering exponential decline of MOS

Fig. 5.6, 5.7 and 5.8 show the MOS versus PLR graphs for cumulative loss of slices. A nice observation of these graphs reveals most of the test-cases lie above 2-3% PLR and thus must have lost more than one type and amount of slices. Fig. 5.12 compares the regression plots of all classes using manual method. It clearly indicates that cumulative loss of slices have greater impact on overall video quality as compared to single slice type loss. Here weights of data points in terms of amount of slices lost within each data point are not considered and all slices are assumed to be of equal importance for generalisation. Moreover, highest gradient in MOS is observed for test-cases in I-class

which means that losing a number of I-slices deteriorates the video quality drastically. B-class maintains the same gradient, however, the regression plot can not be taken totally accountable as for B-class, the number of test-cases are very few. Losing both I and P slices, shows the worst performance. The tail of the graph at higher PLR having better MOS than I and B-class could be explained by the loss of more P-slices than I-slices at those points.

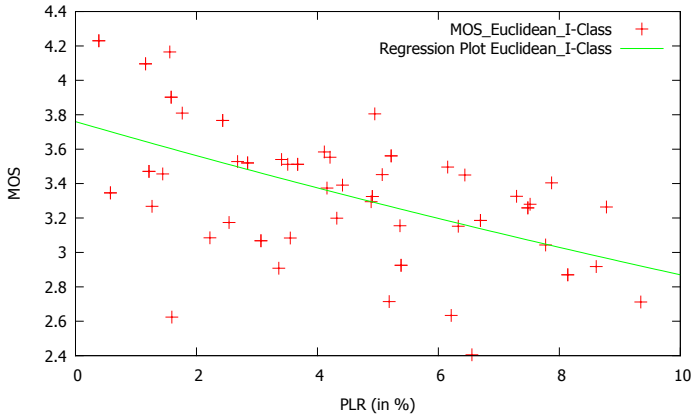


Figure 5.9: Test-cases categorised using Euclidean distance method with loss of data points in I-Class and corresponding PLR vs MOS plot and regression plot considering exponential decline of MOS

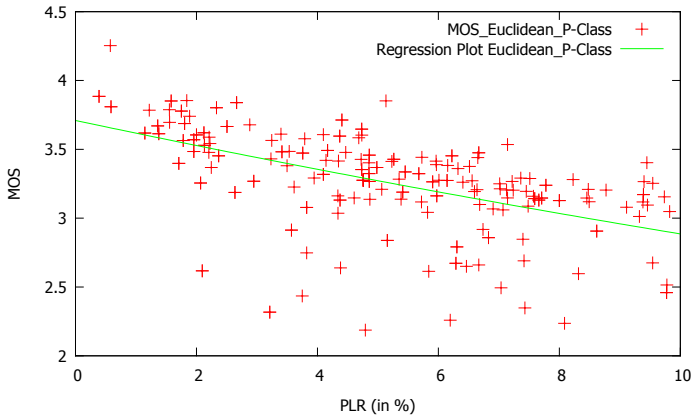


Figure 5.10: Test-cases categorised using Euclidean distance method with loss of data points in P-Class and corresponding PLR vs MOS plot and regression plot considering exponential decline of MOS

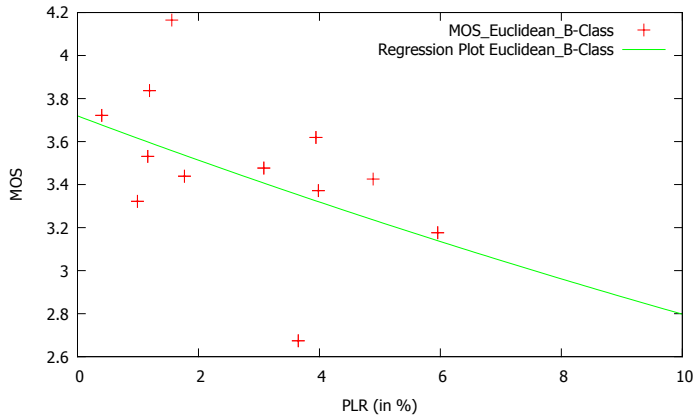


Figure 5.11: Test-cases categorised using Euclidean distance method with loss of data points in B-Class and corresponding PLR vs MOS plot and regression plot considering exponential decline of MOS

In case of data extraction and classification using the *Euclidean* distance method, graphs have been presented in Fig. 5.9, 5.10 and 5.11. Here as well the explanation remains nearly same as they show same characteristics and high correlation with the manual method. Fig. 5.13 shows the regression plots of I, P and B classes on the same graph for comparison. Here the behaviour of plots remains same as the manual process. This cross verifies both the methodologies.An interesting point to note here is that the B-class still performs poorer than the I-class. This performance infers that B-class data points are as important as I-class data points whereas, low gradient of MOS for P-class data points qualifies them as less important than I-class data points.

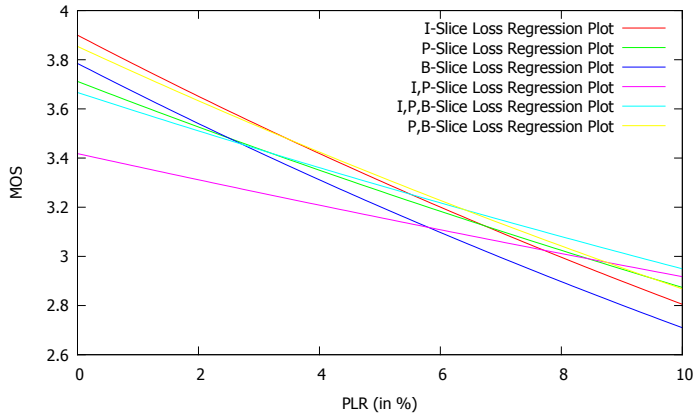


Figure 5.12: Regression plots of all classes categorised using Manual method comparing the impact of specific class on overall video QoE

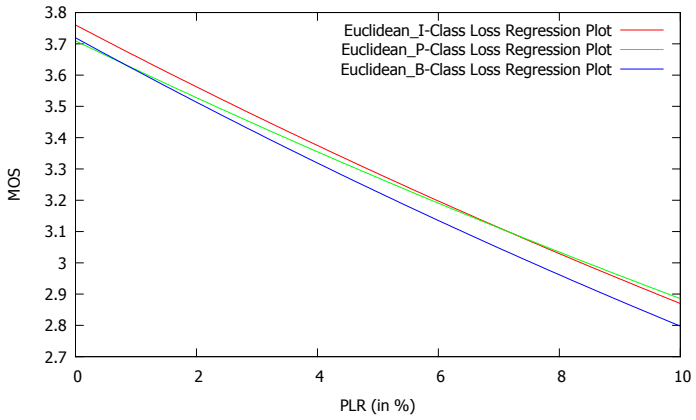


Figure 5.13: Regression plots of all classes categorised using Euclidean distance method comparing the impact of specific class on overall video QoE

5.5 Summary

This chapter presents a brief overview of the difference in impact of loss of specific slice type on the overall QoE. Two different methodologies for classification of test-case, for the purpose of comparative analysis, are presented. Results are presented for analysis in terms of MOS versus PLR plot for each slice type loss. Discussion is present on the impact of each slice type on defining the perceptual quality of video telephony. Both the methodologies are cross verified against each other and they show high correlation with each other and hence verifying the experimental set-up and the results.

There are relevant results to support the idea of prioritization of frame types to provide extra security for enhancement of overall QoE. Hence, the need of an adaptive algorithm to preserve the integrity of the video sequence by protecting some important frames has been verified. Some contrasting results presented in the work show the necessity of further research in the area to define the exact impact of each frame type on the video QoE.

Chapter 6

Conclusion and Future Work

This work has presented a detailed overview of the development, implementation and analysis of an automated framework for the purpose of objectively evaluating the perceptual performance of voice and video over IP products. As discussed in the research questions at the start of the thesis, this research has been able to answer all those questions. Previous chapters have presented a detailed analysis of the products under test in the form of the results relevant to each category. This chapter presents a summary of the work and how this research has answered all the research questions put forward and also the prospective short and long term future work.

6.1 Summary of the Research Work

Section 1.2 has presented an overview of the intended work and posed some research questions to be answered during this work.

- *Can we realise a QoE Assessment Framework from readily available software components that facilitates comprehensive and repeatable studies of VVoIP clients?*

Chapter 3 has presented an overall view of need, design, development and implementation of a framework to assess the perceived call quality in voice and video over IP communications software. It presents two different approaches of testing viz. Intrusive and Non-Intrusive testing and also discusses the pros and cons of each methodology. It presents the architecture of the proposed framework including the components of the product under test and the external tools used for the purpose of framework development. It also presents a sample implementation of the same framework for *IBM*

Sametime, using a plugin as an extension to the client and some external tools. The results and graphs presented for VVoIP service are in accordance to the industry norms and standard guidelines.

High correlation of both testing methodologies affirms and validates the design and working of the framework. The results were also validated using random selection and subjective analysis of voice samples captured. The presented framework answers the research question posed here in terms of presenting a sample implementation and testing of a framework and the results for perceptual analysis of voice and video telephony in terms of the objective metrics defined in industry standards. Dadheech *et al.* (2013a) has been submitted as a result of this chapter.

- *How to analyse the performance of a given video over IP product or specific codec in a given network environment in terms of user perceived quality and perform a comparative analysis of different products?*

Chapter 4 presents a detailed implementation and specifications of the applied framework specific to video phone application software. It explains the tools used for the same and at the centre of the research, it presents the core analysis part of the results generated using the framework. The results are presented for analysis and quality assessment of two different video telephony applications. *IBM Sametime* and *Skype* were used as two different video phone services. Both the applications use different codecs and their implementations and thus this comparative performance analysis also represents the difference between codec implementations.

Results are presented in different standard metrics which are most commonly used for the purpose of performance analysis. However, generalised framework is not based on any given metric or specific tool. Detailed analysis of all the results in form of different metrics has been presented and also the difference in using each metric for a specific functionality is discussed. This chapter has presented the methodology of actually using the framework and analysing the performance of the communications clients under test. Thus, it answers the research question about how to analyse performance of a given product in a given network environment using an appropriate metric. Moreover, comparative performance analysis of two different video phone applications explains an additional use of the framework as a comparative tool. Results presented in terms of the comparison of different codecs or applications, explain the usage of metrics

and their analysis to assess the perceptual quality of video delivered via two different services. This also answers the research question on how to compare two different communications products perceptually. Dadheech *et al.* (2013b) has been accepted for publication as a result of this chapter.

- *How does specific packet/frame type loss affect the overall video QoE?*

Finally chapter 5 presents the idea of how could different parameters other than network impairment, affect the overall video quality. It explains the structure of image encoding in form of pixels, blocks, macro-blocks and slices or frames. It explains the functionality and importance of each type of slice. It presents an applied framework to assess the impact of specific slice loss on video QoE. It presents two different methodologies for classification of test-cases with respect to the types of slices lost during transmission for the purpose of comparative analysis of slice loss patterns on the video QoE. It presents details of the methodologies and sample dataset of categorized test-cases. It then presents the results in the form of different graphs depicting the impact of specific slice loss on each test case and the resultant MOS.

It presents a comparative graph to represent the difference in MOS for same PLR depending on the type of slice lost. This affirms the idea of research initiation to design a methodology and framework for the assessment and comparison of the level of impact of specific slice loss on overall QoE. Comparative results and discussion for different slice types and cumulative slice loss are also presented. This answers the research question on how does different types of slices lost, affect the overall video QoE and how to assess that impact using an applied implementation of the proposed framework.

To conclude, this thesis has answered all the research questions posed and also presented standard results to support and verify the claims. Moreover, during the course of research many other issues have been identified as extensions and key challenges to the presented results which could not be covered in the research. Those issues are presented below next.

6.2 Future Work

During the course of this research and literature review, the core issue identified is, development of a methodology for black box testing for communications products for

perceptual evaluation of their performance. This research presents a solution for the same and also addresses the extensions of the proposed model as short term future work items listed below;

- Extend the present framework for performance evaluation of multi-party and conference voice and video calling support, and analyse the individual or group performance in these types of calls. This involves devising a new model for cumulative MOS evaluation taking care of individual call quality;
- Extend the present framework to feedback the QoE metric estimated using non-intrusive methodology back to the client. Purpose of this feedback system is to develop a training model within the application client to learn the occurrence of low QoE score and adapt its application or codec implementation parameters for optimization of available resources and maximising the QoE for multi-party calls;
- Propose an external solution for intelligent packetization of image slices into IP packets depending on their level of importance assessed by the presented results to prevent the loss of some specific slices for maximising the overall video QoE in a network loss environment.

Moreover, there have been some issues identified as a complementary research question to be answered. These issues involve detailed literature review, major modifications to the framework and adapting different methodologies to answer the research questions. These issues have been considered for long term future work and are listed below;

- Extension to Opinion model to include more parameters affecting the end-to-end video quality viz. resolution, key frame interval.
- Identify the impact of video content on the overall video quality in an impaired network environment and to devise generalised algorithms for all video contents for protection of important image frames and maximising video QoE.
- Propose major changes in Opinion model to encapsulate the effect of slice type and video content on the estimated MOS and to develop generalised variable set for video codecs.

- Discover an optimized variable set for opinion model for different codecs based on the implementation i.e. profile and level used.

References

(2013). Google hangout. http://www.google.com/hangouts/. 30

(2013). IBM Sametime Unified Telephony. http://www.ibm.com/. 68, 71

(2013). Skype. http://www.skype.com. 30, 68, 71

(2013). Video Test Media [derf's collection]. http://media.xiph.org/video/derf/. 59

Agboma, F. & Liotta, A. (2008). QoE-aware QoS management. In *Proceedings of the 6th International Conference on Advances in Mobile Computing and Multimedia*, 111–116, ACM. 39

Atenas, M., Garcia, M., Canovas, A. & Lloret, J. (2010). A MPEG-2/MPEG-4 quantizer to improve the video quality in IPTV services. In *Networking and Services (ICNS), 2010 Sixth International Conference on*, 49–54, IEEE. 14

Bankoski, J. (2011). Intro to webm. In *Proceedings of the 21st international workshop on Network and operating systems support for digital audio and video*, 1–2, ACM. 28

Bankoski, J., Wilkins, P. & Xu, Y. (2011). Technical overview of vp8, an open source video codec for the web. In *Multimedia and Expo (ICME), 2011 IEEE International Conference on*, 1–6, IEEE. 29

Brooks, P. & Hestnes, B. (2010). User measures of quality of experience: Why being objective and quantitative is important. *Network, IEEE*, **24**, 8–13. 18

Calyam, P., Ekici, E., Lee, C.G., Haffner, M. & Howes, N. (2007). A GAP-Model based framework for online VVoIP QoE measurement. *Communications and Networks, Journal of*, **9**, 446–456. 42

Carbone, M. & Rizzo, L. (2010). Dummynet revisited. *ACM SIGCOMM Computer Communication Review*, **40**, 12–20. 54

Chen, J.W., Kao, C.Y. & Lin, Y.L. (2006). Introduction to H.264 advanced video coding. In *Proceedings of the 2006 Asia and South Pacific Design Automation Conference*, 736–741, IEEE Press. 28

Chen, K.T., Tu, C.C. & Xiao, W.C. (2009). OneClick: A framework for measuring network quality of experience. In *INFOCOM 2009, IEEE*, 702–710, IEEE. 41

Cherif, W., Ksentini, A., Négru, D. & Sidibe, M. (2012). A_PSQA: PESQ-like non-intrusive tool for QoE prediction in VoIP services. In *Communications (ICC), 2012 IEEE International Conference on*, 2124–2128, IEEE. 41

CISCO (2001). Quality of Service for Voice over IP. Tech. rep., Cisco. 4

Cole, R.G. & Rosenbluth, J.H. (2001). Voice over IP performance monitoring. *ACM SIGCOMM Computer Communication Review*, **31**, 9–24. 21

Combs, G. *et al.* (2007). Wireshark. *Web page: http://www. wireshark. org/last modified*, 12–02. 55

Conway, A.E. (2002). A passive method for monitoring Voice-over-IP call quality with ITU-T objective speech quality measurement methods. In *Communications, 2002. ICC 2002. IEEE International Conference on*, vol. 4, 2583–2586, IEEE. 42

Cui, H., Tang, K. & Cheng, T. (1998). Audio as a support to low bit rate multimedia communication. In *Communication Technology Proceedings, 1998. ICCT'98. 1998 International Conference on*, 544–547, IEEE. 26

da Silva, A.P.C., Varela, M., de Souza e Silva, E., Leão, R.M. & Rubino, G. (2008). Quality assessment of interactive voice applications. *Computer Networks*, **52**, 1179–1192. 40

Dadheech, H., Han, Y., Jennings, B., Malone, D., Murphy, L., Dunne, J. & O'Sullivan, P. (2013a). A Quality-of-Experience assessment framework for management of enterprise Voice/Video-over-IP services. *Communications Magazine, IEEE*, **Under Review**. 9, 96

Dadheech, H., Jennings, B. & Dunne, J. (2013b). A call quality assessment and analysis framework for video telephony applications in enterprise networks. In *Global Information Infrastructure and Networking Symposium (GIIS), 2013*, Accepted and to be published in IEE EXPLORE. 9, 97

Dai, Q. & Lehnert, R. (2010). Impact of packet loss on the perceived video quality. In *Evolving Internet (INTERNET), 2010 Second International Conference on*, 206–209, IEEE. 43

De Rango, F., Tropea, M., Fazio, P. & Marano, S. (2006). Overview on VoIP: subjective and objective measurement methods. *International Journal of Computer Science and Network Security*, **6**, 140–153. 42, 58

De Simone, F., Naccari, M., Tagliasacchi, M., Dufaux, F., Tubaro, S. & Ebrahimi, T. (2009). Subjective assessment of H.264/AVC video sequences transmitted over a noisy channel. In *Quality of Multimedia Experience, 2009. QoMEx 2009. International Workshop on*, 204–209, IEEE. 39

Ding, L. & Goubran, R.A. (2003). Assessment of effects of packet loss on speech quality in VoIP. In *Haptic, Audio and Visual Environments and Their Applications, 2003. HAVE 2003. Proceedings. The 2nd IEEE Internatioal Workshop on*, 49–54, IEEE. 26

Ding, L., Radwan, A., El-Hennawey, M.S. & Goubran, R. (2007). Performance study of objective voice quality measures in VoIP. In *Computers and Communications, 2007. ISCC 2007. 12th IEEE Symposium on*, 197–202, IEEE. 13

Engelke, U. & Zepernick, H.J. (2007). Perceptual-based quality metrics for image and video services: A survey. In *Next Generation Internet Networks, 3rd EuroNGI Conference on*, 190–197, IEEE. 40

Fathi, H., Chakraborty, S.S. & Prasad, R. (2006). Optimization of SIP session setup delay for VoIP in 3G wireless networks. *Mobile Computing, IEEE Transactions on*, **5**, 1121–1132. 25

Fiedler, M., Hossfeld, T. & Tran-Gia, P. (2010). A generic quantitative relationship between quality of experience and quality of service. *Network, IEEE*, **24**, 36–41. 38

Goode, B. (2002). Voice over internet protocol (VoIP). *Proceedings of the IEEE*, **90**, 1495–1517. 1

Greengrass, J., Evans, J. & Begen, A.C. (2009a). Not all packets are equal, part 1: Streaming video coding and SLA requirements. *Internet Computing, IEEE*, **13**, 70–75. viii, 16, 32, 34, 43, 79

Greengrass, J., Evans, J. & Begen, A.C. (2009b). Not all packets are equal, part 2: The impact of network packet loss on video quality. *Internet Computing, IEEE*, **13**, 74–82. 43

Hartigan, J.A. & Wong, M.A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **28**, 100–108. 83, 85

Hoeher, T., Petraschek, M., Tomic, S. & Hirschbichler, M. (2007). Evaluating performance characteristics of SIP over IPv6. *Journal of Networks*, **2**, 40–50. 25

Hohlfeld, O., Geib, R. & Hasslinger, G. (2008). Packet loss in real-time services: Markovian models generating QoE impairments. In *Quality of Service, 2008. IWQoS 2008. 16th International Workshop on*, 239–248, IEEE. 44

Horak, R. (2007). *Telecommunications and data communications handbook*. Wiley-Interscience. 25

ISO & IEC (1988). Moving picture experts group. 3

ITU-T (1994). G.113: Transmission systems and media general recommendations on the transmission quality for an entire international telephone connection: Transmission impairments. *International Telecommunication Union*. 22

Joskowicz, J. & Ardao, J. (2009). Enhancements to the opinion model for videotelephony applications. In *Proceedings of the 5th International Latin American Networking Conference*, 87–94, ACM. 45

Joskowicz, J., Ardao, J.C.L. & Sotelo, R. (2011). Including the effects of video content in the ITU-T G.1070 video quality function. *Cadernos de Informática*, **6**, 227–232. 42, 45, 58, 69, 76

Jumisko-Pyykkö, S. & Häkkinen, J. (2005). Evaluation of subjective video quality of mobile devices. In *Proceedings of the 13th annual ACM international conference on Multimedia*, 535–538, ACM. 18

Keepence, B. (1999). Quality of service for voice over IP. *IET*. 38

Kent, R. & Tepper, H. (2005). Enterprise video conferencing: Ready for prime time. *Green Spring Partners*. 3

Kim, S. & Yoon, Y. (2008). Video customization system using MPEG standards. In *Multimedia and Ubiquitous Engineering, 2008. MUE 2008. International Conference on*, 475–480, IEEE. 14, 38

Koumaras, H., Kourtis, A. & Martakos, D. (2005). Evaluation of video quality based on objectively estimated metric. *Communications and Networks, Journal of*, **7**, 235–242. 40

Kuipers, F., Kooij, R., De Vleeschauwer, D. & Brunnström, K. (2010). Techniques for measuring quality of experience. In *Wired/wireless internet communications*, 216–227, Springer. 17, 39

Kwon, S.k., Tamhankar, A. & Rao, K. (2006). Overview of H.264/MPEG-4 part 10. *Journal of Visual Communication and Image Representation*, **17**, 186–216. 16

Lambrinos, L. & Kirstein, P. (2007). Integrating Voice over IP services in IPv4 and IPv6 networks. In *Computing in the Global Information Technology, 2007. ICCGI 2007. International Multi-Conference on*, 54–54, IEEE. 25

Le Callet, P., Möller, S. & Perkis, A. (2012). Qualinet white paper on definitions of quality of experience (2012). 6

Lee, A. (2007). VirtualDub. *web site: www.virtualdub.org CS MSU GRAPHICS&MEDIA LAB ABOUT VIDEO GROUP*. 56

Lu, X., Tao, S., Zarki, M. & Guérin, R. (2003). Quality-based adaptive video over the Internet. In *Proceedings of CNDS*, Citeseer. 41

Luigi, R. (1999). The DummyNet Project. http://info.iet.unipi.it/ luigi/dummynet/. 54, 59, 69, 82

ManyCam, L. (2013). ManyCam. http://www.manycam.com/. 57

MASUDA, M. & HAYASHI, T. (2006). Non-intrusive quality monitoring method of VoIP speech based on network performance metrics. *IEICE transactions on communications*, **89**, 304–312. 18

McFarland, M.A., Pinson, M.H. & Wolf, S. (2007). Batch Video Quality Metric (BVQM) Users Manual. *Insititute for Telecommunication Sciences, Boulder, Colorado, USA, http://www. its. bldrdoc. gov/pub/ntia-rpt/06-441a*. 55, 70

Menth, M., Binzenhöfer, A. & Mühleck, S. (2009). Source models for speech traffic revisited. *IEEE/ACM Transactions on Networking (TON)*, **17**, 1042–1051. 26

Moorthy, A.K., Seshadrinathan, K., Soundararajan, R. & Bovik, A.C. (2010). Wireless video quality assessment: A study of subjective scores and objective algorithms. *Circuits and Systems for Video Technology, IEEE Transactions on*, **20**, 587–599. 18

Mu, M., Gostner, R., Mauthe, A., Tyson, G. & Garcia, F. (2009). Visibility of individual packet loss on H.264 encoded video stream: A user study on the impact of packet loss on perceived video quality. In *IS&T/SPIE Electronic Imaging*, 725302–725302, International Society for Optics and Photonics. 43

Muppala, J.K., Bancherdvanich, T. & Tyagi, A. (2000). VoIP performance on differentiated services enabled network. In *Networks, 2000.(ICON 2000). Proceedings. IEEE International Conference on*, 419–423, IEEE. 26

Nemethova, O., Ries, M., Zavodsky, M. & Rupp, M. (2006). PSNR-based estimation of subjective time-variant video quality for mobiles. *Proc. of the MESAQUIN*. 41

Nguyen, T. & Zakhor, A. (2002). Distributed video streaming with forward error correction. In *Packet Video Workshop*, vol. 2002. 4

Ohm, J.R. (2004). *Multimedia communication technology: Representation, transmission and identification of multimedia signals*. Springer Verlag. 75

Pérez, P., Macías, J., Ruiz, J.J. & García, N. (2011). Effect of packet loss in video quality of experience. *Bell Labs Technical Journal*, **16**, 91–104. 44

Pinson, M. & Wolf, S. (2005). *Reduced reference video calibration algorithms*. National Telecommunications and Information Administration. 56

Pinson, M.H. & Wolf, S. (2003). Comparing subjective video quality testing methodologies. In *Visual Communications and Image Processing 2003*, 573–582, International Society for Optics and Photonics. 39

Pinson, M.H. & Wolf, S. (2004). A new standardized method for objectively measuring video quality. *Broadcasting, IEEE Transactions on*, **50**, 312–322. 66

Puri, A. & Eleftheriadis, A. (1998). MPEG-4: An object-based multimedia coding standard supporting mobile applications. *Mobile Networks and Applications*, **3**, 5–32. 27

Ramakrishnan, R. & Kumar, P. (2008). Performance analysis of different codecs in VoIP using SIP. *The Conference on Mobile and Pervasive Computing*, 142–145. 12

Rec, I. (1988). G.711: Pulse code modulation (PCM) of voice frequencies. *International Telecommunication Union, Geneva*. 26

Rec, I. (2001). 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. *International Telecommunication Union*. 19

Rec, I. (2005). H.264, Advanced video coding for generic audiovisual services. *ITU-T Rec. H. 264-ISO/IEC 14496-10 AVC*. 35

Rec, I. (2007). G.1070, Opinion model for videotelephony applications. *International Telecommunication Union*. 23, 58, 65, 66, 69

Rec, I. (2008). P.910, Subjective video quality assessment methods for multimedia applications". *International Telecommunication Union*. 68

Rec, I. (2009). BT.500-11 (2002): Methodology for the subjective assessment of the quality of television pictures. *International Telecommunication Union*. 39

Rec, I.T. (2003). G.107-the e model, a computational model for use in transmission planning. *International Telecommunication Union*. viii, 20, 21, 58

Rec, I.T. (2011). G.107.1- Wideband E-Model. *International Telecommunication Union*. 20

Reibman, A.R., Sen, S. & Van der Merwe, J. (2004). Network monitoring for video quality over IP. In *Picture Coding Symposium*. 41

Rix, A.W. (2003). Comparison between subjective listening quality and P.862 PESQ score. *Proc. Measurement of Speech and Audio Quality in Networks (MESAQIN03), Prague, Czech Republic*. 19

Rix, A.W., Beerends, J.G., Hollier, M.P. & Hekstra, A.P. (2001). Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, vol. 2, 749–752, IEEE. 19

Rix, A.W., Beerends, J.G., Kim, D.S., Kroon, P. & Ghitza, O. (2006). Objective assessment of speech and audio qualitytechnology and applications. *Audio, Speech, and Language Processing, IEEE Transactions on*, **14**, 1890–1901. 40

Rosenberg, J., Schulzrinne, H., Camarillo, G., Johnston, A., Peterson, J., Sparks, R., Handley, M., Schooler, E. *et al.* (2002). SIP: session initiation protocol. Tech. rep., RFC 3261, Internet Engineering Task Force. 25

Roychoudhuri, L., Al-Shaer, E., Hamed, H. & Brewster, G.B. (2003). Audio transmission over the Internet: Experiments and observations. In *Communications, 2003. ICC'03. IEEE International Conference on*, vol. 1, 552–556, IEEE. 26

Salah, K. (2008). Deploying VoIP in Existing IP Networks. *VoIP Handbook: Applications, Technologies, Reliability, and Security*, 3. 12

Savolaine, C. (2001). QoS/VoIP overview. In *IEEE Communications Quality & Reliability (CQR 2001) International Workshop*, vol. 4. 38

Schierl, T., Gruneberg, K. & Wiegand, T. (2009). Scalable video coding over RTP and MPEG-2 transport stream in broadcast and IPTV channels. *Wireless Communications, IEEE*, **16**, 64–71. 14

Schwarz, H., Marpe, D. & Wiegand, T. (2007). Overview of the scalable video coding extension of the H.264/AVC standard. *Circuits and Systems for Video Technology, IEEE Transactions on*, **17**, 1103–1120. 4

Serral-Gracià, R., Cerqueira, E., Curado, M., Yannuzzi, M., Monteiro, E. & Masip-Bruin, X. (2010). An overview of quality of experience measurement challenges for video applications in IP networks. In *Wired/Wireless Internet Communications*, 252–263, Springer. 66

Seshadrinathan, K., Soundararajan, R., Bovik, A.C. & Cormack, L.K. (2010a). Study of subjective and objective quality assessment of video. *Image Processing, IEEE Transactions on*, **19**, 1427–1441. 5, 39

Seshadrinathan, K., Soundararajan, R., Bovik, A.C. & Cormack, L.K. (2010b). A subjective study to evaluate video quality assessment algorithms. In *IS&T/SPIE Electronic Imaging*, 75270H–75270H, International Society for Optics and Photonics. 40

Sims, P.J. (2007). A study on Video over IP and the Effects on FTTx architectures. In *Globecom Workshops, 2007 IEEE*, 1–4, IEEE. 16

Staelens, N., Moens, S., Van den Broeck, W., Marien, I., Vermeulen, B., Lambert, P., Van de Walle, R. & Demeester, P. (2010). Assessing quality of experience of IPTV and video on demand services in real-life environments. *Broadcasting, IEEE Transactions on*, **56**, 458–466. 40

Sullivan, G.J., Topiwala, P.N. & Luthra, A. (2004). The H.264/AVC advanced video coding standard: Overview and introduction to the fidelity range extensions. In *Optical Science and Technology, the SPIE 49th Annual Meeting*, 454–474, International Society for Optics and Photonics. 28

Takahashi, A., Yoshino, H. & Kitawaki, N. (2004). Perceptual QoS assessment technologies for VoIP. *Communications Magazine, IEEE*, **42**, 28–34. 38, 48

Takahashi, A., Kurashima, A. & Yoshino, H. (2006). Objective assessment methodology for estimating conversational quality in VoIP. *Audio, Speech, and Language Processing, IEEE Transactions on*, **14**, 1984–1993. 22

Takahashi, A., Hands, D. & Barriac, V. (2008). Standardization activities in the ITU for a QoE assessment of IPTV. *Communications Magazine, IEEE*, **46**, 78–84. 40

Tasaka, S. & Misaki, N. (2009). Maximizing QoE of interactive services with audio-video transmission over bandwidth guaranteed IP networks. In *Global Telecommunications Conference, 2009. GLOBECOM 2009. IEEE*, 1–7, IEEE. 42

Tasaka, S., Yoshimi, H., Hirashima, A. & Nunome, T. (2008). The effectiveness of a QoE-based video output scheme for audio-video IP transmission. In *Proceedings of the 16th ACM international conference on Multimedia*, 259–268, ACM. 42

Van Moorsel, A. (2001). Metrics for the internet age: Quality of experience and quality of business. In *Fifth International Workshop on Performability Modeling of Computer and Communication Systems, Arbeitsberichte des Instituts für Informatik, Universität Erlangen-Nürnberg, Germany*, vol. 34, 26–31, Citeseer. 6, 38

Varga, I., Proust, S. & Taddei, H. (2009). ITU-T G.729.1 scalable codec for new wideband services. *Communications Magazine, IEEE*, **47**, 131–137. 26

Varshney, U., Snow, A., McGivern, M. & Howard, C. (2002). Voice over IP. *Communications of the ACM*, **45**, 89–96. 1

Vatolin, D., Moskvin, A., Petrov, O. & Trunichkin, N. (2009). MSU video quality measurement tool. 70

Venkataraman, M. & Chatterjee, M. (2011). Inferring video QoE in real time. *Network, IEEE*, **25**, 4–13. 41

Wang, Y. (2006). Survey of objective video quality measurements. *EMC Corporation Hopkinton, MA*, **1748**. 39

Wang, Z. & Li, Q. (2007). Video quality assessment using a statistical model of human visual speed perception. *JOSA A*, **24**, B61–B69. 23

Wang, Z., Bovik, A.C., Sheikh, H.R. & Simoncelli, E.P. (2003a). The SSIM index for image quality assessment. *MATLAB implementation available online from: http://www. cns. nyu. edu/˜ lcv/ssim*. 23, 66

Wang, Z., Sheikh, H.R. & Bovik, A.C. (2003b). Objective video quality assessment. *The handbook of video databases: design and applications*, 1041–1078. 40

Wang, Z., Bovik, A.C., Sheikh, H.R. & Simoncelli, E.P. (2004). Image quality assessment: From error visibility to structural similarity. *Image Processing, IEEE Transactions on*, **13**, 600–612. 23

Wenger, S., Hannuksela, M., Stockhammer, T., Westerlund, M. & Singer, D. (2005). RFC 3984; RTP payload format for H.264 video. *IETF, February*. viii, 31, 35

Wiegand, T., Sullivan, G.J., Bjontegaard, G. & Luthra, A. (2003). Overview of the H.264/AVC video coding standard. *Circuits and Systems for Video Technology, IEEE Transactions on*, **13**, 560–576. 3

Winkler, S. (2009). Video quality measurement standards: Current status and trends. In *Information, Communications and Signal Processing, 2009. ICICS 2009. 7th International Conference on*, 1–5, IEEE. 18, 39

Winkler, S. & Mohandas, P. (2008). The evolution of video quality measurement: From PSNR to hybrid metrics. *Broadcasting, IEEE Transactions on*, **54**, 660–668. 18, 40, 65, 66

Wolf, S. (2009). A Full Reference (FR) method using causality processing for estimating variable video delays. Tech. rep., NTIA Technical Memorandum TM-10-463. 56

Wolf, S. & Pinson, M. (2002). Video quality measurement techniques. *2002.*. 55

Yamada, T., Miyamoto, Y. & Serizawa, M. (2007). No-reference video quality estimation based on error-concealment effectiveness. In *Packet Video 2007*, 288–293, IEEE. 42

Yamagishi, K. & Hayashi, T. (2008). Parametric packet-layer model for monitoring video quality of IPTV services. In *Communications, 2008. ICC'08. IEEE International Conference on*, 110–114, IEEE. 45

Yammine, G., Wige, E. & Kaup, A. (2010). Blind GOP structure analysis of MPEG-2 and H.264/AVC decoded video. In *Picture Coding Symposium (PCS), 2010*, 258–261, IEEE. 45

You, J., Reiter, U., Hannuksela, M.M., Gabbouj, M. & Perkis, A. (2010). Perceptual-based quality assessment for audio–visual services: A survey. *Signal Processing: Image Communication*, **25**, 482–501. 39

Zinner, T., Hohlfeld, O., Abboud, O. & Hossfeld, T. (2010). Impact of frame rate and resolution on objective QoE metrics. In *Quality of Multimedia Experience (QoMEX), 2010 Second International Workshop on*, 29–34, IEEE. 42, 44, 48