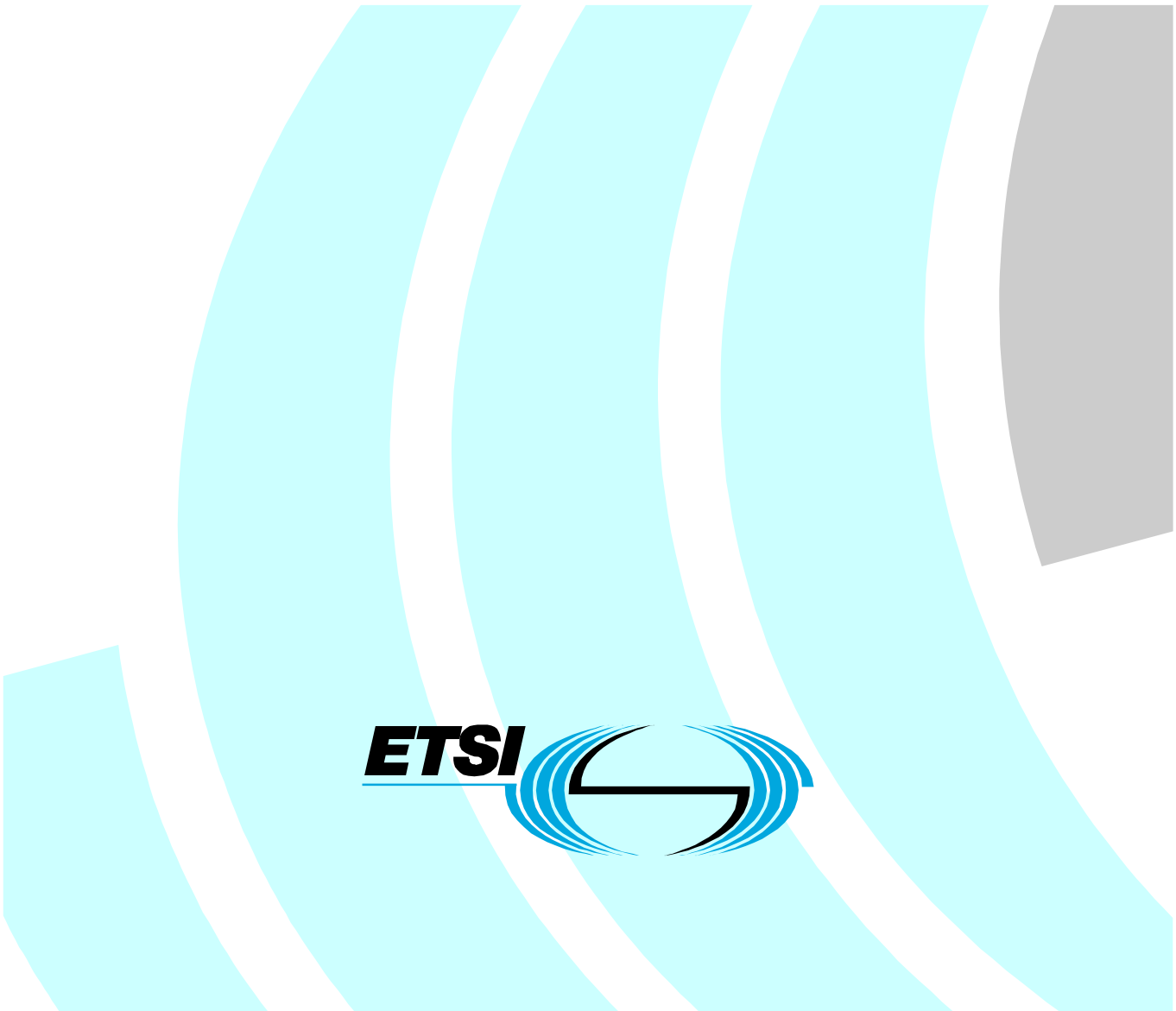
ETSI TR 102 479 V1.1.1 (2006-02)

Technical Report

Telecommunications and Internet converged Services and Protocols for Advanced Networking (TISPAN); Review of available material on QoS requirements of Multimedia Services



Reference
DTR/TISPAN-05006-Tech

Keywords
multimedia, QoS

ETSI

650 Route des Lucioles
F-06921 Sophia Antipolis Cedex - FRANCE

Tel.: +33 4 92 94 42 00 Fax: +33 4 93 65 47 16

Siret N° 348 623 562 00017 - NAF 742 C
Association à but non lucratif enregistrée à la
Sous-Préfecture de Grasse (06) N° 7803/88

Important notice

Individual copies of the present document can be downloaded from:
http://www.etsi.org

The present document may be made available in more than one electronic version or in print. In any case of existing or perceived difference in contents between such versions, the reference version is the Portable Document Format (PDF). In case of dispute, the reference shall be the printing on ETSI printers of the PDF version kept on a specific network drive within ETSI Secretariat.

Users of the present document should be aware that the document may be subject to revision or change of status. Information on the current status of this and other ETSI documents is available at
http://portal.etsi.org/tb/status/status.asp

If you find errors in the present document, please send your comment to one of the following services:
http://portal.etsi.org/chaircor/ETSI_support.asp

Copyright Notification

Contents

Intellectual Property Rights

IPRs essential or potentially essential to the present document may have been declared to ETSI. The information pertaining to these essential IPRs, if any, is publicly available for **ETSI members and non-members**, and can be found in ETSI SR 000 314: *"Intellectual Property Rights (IPRs); Essential, or potentially Essential, IPRs notified to ETSI in respect of ETSI standards"*, which is available from the ETSI Secretariat. Latest updates are available on the ETSI Web server (http://webapp.etsi.org/IPR/home.asp).

Pursuant to the ETSI IPR Policy, no investigation, including IPR searches, has been carried out by ETSI. No guarantee can be given as to the existence of other IPRs not referenced in ETSI SR 000 314 (or the updates on the ETSI Web server) which are, or may be, or may become, essential to the present document.

Foreword

This Technical Report (TR) has been produced by ETSI Technical Committee Telecommunications and Internet converged Services and Protocols for Advanced Networking (TISPAN).

1 Scope

The present document provides an overview of factors that influence user perceived quality in TISPAN compliant systems supporting multimedia applications.

Multimedia applications are defined as those which combine different media types with potentially fundamentally different properties and inter-relationships. Examples of media types are audio, video, animation, still pictures, graphics and data (text). Multimedia applications include videoconferencing, audio streaming, CCTV, broadcast TV, etc. Although part of an integrated application, media flows within multimedia applications may be very different in terms of transmission quality requirements.

The present document defines the audio and video quality requirements for a variety of multimedia applications involving conversational and streaming media flows and the transmission quality requirements to support these in TISPAN systems. Video applications are restricted to those involving screens of medium size (12') and upwards.

A classification system is included to describe the quality aspects of multimedia systems, their media components and the transmission quality requirements in TISPAN systems.

Issues of Media synchronization are also included.

2 References

For the purposes of this Technical Report (TR), the following references apply:

[1] ETSI TR 102 274 (V1.1.1): "Human Factors (HF); Guidelines for real-time person-to-person communication services".

[2] ITU-T Recommendation E.800: "Terms and definitions related to quality of service and network performance including dependability".

[3] ITU-T Recommendation G.1000: "Communications Quality of Service: A framework and definitions".

[4] ETSI TS 185 001 (V1.1.1): "Telecommunications and Internet converged Services and Protocols for Advanced Networking (TISPAN); Next Generation Network (NGN); Quality of Service (QoS) Framework and Requirements".

[5] ETSI ES 282 003: "Telecommunications and Internet converged Services and Protocols for Advanced Networking (TISPAN); Resource and Admission Control Sub-system (RACS); Functional Architecture".

[6] ITU-T Recommendation P.911 (1998): "Subjective audiovisual quality assessment methods for multimedia applications".

[7] ITU-T Recommendation G.1010 (2001): "End-user multimedia QoS categories".

[8] ITU-R Recommendation BT.601: "Studio encoding parameters of digital television for standard 4:3 and wide-screen 16:9 aspect ratios".

[9] ITU-T Recommendation Y.1541 (2002): "Network performance objectives for IP-based services".

[10] ITU-T Recommendation Y.1540: "Internet protocol data communication service - IP packet transfer and availability performance parameters".

[11] ITU-T Recommendation H.360: "An architecture for end-to-end QoS control and signalling".

[12] ITU-T Recommendation G.1050 (2005): "Network model for evaluating multimedia transmission performance over internet protocol".

[13] ETSI TS 123 107 (V6.3.0): "Digital cellular telecommunications system (Phase 2+); Universal Mobile Telecommunications System (UMTS); Quality of Service (QoS) concept and architecture (3GPP TS 23.107 version 6.3.0 Release 6)".

[14] ANSI T1.522 (2000): "Quality of Service for Business Multimedia Conferencing".

[15] ITU-T Recommendation G.711: "Pulse Code Modulation (PCM) of voice frequencies".

[16] ITU-T Recommendation G.722: "7 kHz audio-coding within 64 kbit/s".

[17] IETF RFC 3951: "Internet Low Bit Rate Codec (iLBC)".

[18] S. V. Andersen, W. B. Kleijn, S. Hagen, J. Linden, M. N. Murthi and J. ILBC Skoglund: "A linear predictive coder with robustness to packet losses". 2002 IEEE Workshop on Speech Coding, Tsukuba, Ibaraki, Japan, 6-9 October, 2002.

[19] ITU-T Recommendation G.729: "Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear-prediction (CS-ACELP)".

[20] IETF RFC 3551: "RTP Profile for Audio and Video Conferences with Minimal Control".

[21] IETF RFC 3550: "RTP: A transport Protocol for Real-Time Applications".

[22] ITU-T Recommendation G.723.1: "Speech coders: Dual rate speech coder for multimedia communications transmitting at 5.3 and 6.3 kbit/s".

[23] TIA/EIA-TSB-116: "Telecommunications - IP Telephony - Voice Quality Recommendations for IP Telephony".

[24] ITU-T Recommendation G.114: "One-way transmission time".

[25] Y. J. Liang, N. Färber and B. Girod: "Adaptive Playout Scheduling and Loss Concealment for Voice Communication Over IP Networks". IEEE Transactions on Multimedia, vol. 5, No. 4 (1993).

[26] ITU-T Recommendation G.107: "The E-model, a computational model for use in transmission planning".

[27] ETSI TR 101 329-6: "Telecommunications and Internet Protocol Harmonization Over Networks (TIPHON) Release 3; End-to-end Quality of Service in TIPHON systems; Part 6: Actual measurements of network and terminal characteristics and performance parameters in TIPHON networks and their influence on voice quality".

[28] A. Clark: "Modelling the effects of Burst Packet Loss and the Recency on Subjective Voice Quality". The 3rd IP Telephony Workshop 2002, New York 28 April-2 May 2002.

[29] W. Jiang and H. Schulzrinne: "Comparison and Optimization of Packet Loss Repair Methods on VoIP Perceived Quality under Bursty Loss". NOSSDAV"02, Miami Beach, 12-14 May 2002.

[30] ETSI TS 101 329-5 (V1.1.1): "Telecommunications and Internet protocol Harmonization Over Networks (TIPHON) Release 3; Technology Compliance Specification; Part 5: Quality of Service (QoS) measurement methodologies".

[31] A. Duric: "Speech/Audio coding for IP networks. ETSI Speech Processing, Transmission and Quality Aspects (STQ)". Workshop Compensating for Packet Loss in Real-Time Applications, 11th February 2003.

[32] ITU-T Recommendation G.722.1: "Low-complexity coding at 24 and 32 kbit/s for hands-free operation in systems with low frame loss".

[33] ETSI ETR 250: "Transmission and Multiplexing (TM); Speech communication quality from mouth to ear for 3,1 kHz handset telephony across networks".

[34] ETSI TR 102 356 (V1.1.1): "Speech Processing, Transmission and Quality Aspects (STQ); Application and enhancements of the E-Model (ETR 250); Overview of available documentation and ongoing work".

[35] ISO/IEC 11172 (MPEG 1, 5 parts): "Information technology - Coding of moving pictures and associated audio for digital storage media at up to about 1,5 Mbit/s".

[36] ISO/IEC 13818 (MPEG 2, 11 parts): "Information technology - Generic coding of moving pictures and associated audio information".

[37] ISO/IEC 14496 (MPEG 4; currently in 16 parts): "Information technology - Coding of audio-visual objects".

[38] Fraunhofer Institute for Integrated Circuits IIS: "MPEG-4 AAC-LD Low-Delay High-Quality Audio Coding".

[39] ISO/IEC 14496-3 Amd1 (2003): "Bandwidth Extension".

[40] K. Brandenburg: "MP3 and AAC explained". AES 17th International Conference on High Quality Coding, 1999.

[41] ITU-R Recommendation BS.1387: "Method for objective measurements of perceived audio quality".

[42] ITU-T Recommendation H.261: "Video codec for audiovisual services at p x 64 kbit/s".

[43] ITU-T Recommendation H.263: "Video coding for low bit rate communication".

[44] ITU-T Recommendation H.264: "Advanced video coding for generic audiovisual services (Also known as ISO/IEC 14496-10)".

[45] R. Koenen: "MPEG 4 multimedia for our time". IEEE Spectrum, vol 36, No. 2, February 1999, pp. 26-33.

[46] ACTS AC 314: "Vis-à-Vis Final report". Reference number A0314/NSSL/PB/DR/P/005/b1, 30 June 1999.

[47] Yadavalli, G,. Masry, M. and Hemami: "Frame rate preferences in low bit rate video". IEEE International Conference on Image Processing (ICIP 2003) Barcelona, September, 14-17 2003.

[48] ITU-R Recommendation BT.500: "Methodology for the subjective Assessment of the Quality of Television Pictures".

[49] ITU-T Recommendation H.320: "Narrow-band visual telephone systems and terminal equipment".

[50] ITU-T Recommendation H.323: "Packet-based multimedia communications systems".

[51] G. Côté, B. Erol, M. Gallant and F. Kossentini: "H.263+: Video coding at low bit rates". IEEE Transactions on Circuits and Systems for Video Technology, vol. 8, No. 7, November 1998, pp. 849-866.

[52] T. Wiegand, G. J. Sullivan, G. Bjøntegaard and A. Luthra: "Overview of the H.264/AVC Video Coding Standard". IEEE Transactions on Circuits and Systems for Video Technology, vol. 13, No. 7, July 2003, pp. 560-576.

[53] S. Wenger: "H.264/AVC over IP". IEEE Transactions on Circuits and Systems for Video Technology,vol. 13, No. 7, July 2003, pp. 645-656.

[54] SMPTE 421M.: "Draft SMPTE Standard for Television - VC-1 Compressed Video Bitstream Format and Decoding Process", 2005.

[55] J. Bennett and A. Bock.: "In-Depth Review of Advanced Coding Technologies for Low Bit Rate Broadcast Applications". IBC 2003, Amsterdam, 11-15 September 2003.

[56] R. T. Apteker, J. A. Fisher, V. S. Kisimov and H. Neishlos: "Video acceptability and Frame Rate". IEEE Multimedia, vol. 2, No. 3, 1995, pp. 32-40.

[57] Video Quality Experts Group (VQEG): "Final report from the Video Quality Experts Group on the validation of objective models of video quality assessment". March 2000. http://www.its.bldrdoc.gov/vqeg/projects/frtv_phaseI/index.php.

[58] Video Quality Experts Group (VQEG): "Final report from the Video Quality Experts Group on the validation of objective models of video quality assessment, Phase II". August 2003. http://www.its.bldrdoc.gov/vqeg/projects/frtv_phaseII/index.php.

[59] ITU-R Recommendation BT.1683 (2004): "Objective perceptual video quality measurement techniques for standard definition digital broadcast television in the presence of a full reference".

[60] ITU-T Recommendation J.144 (2004): "Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference".

[61] H. Thoma: "Delay in Video Transmission and Rate control". http://www.hthoma.de/video/delay/.

[62] T. Hayashi, S. Yamasaki, N. Morita, H. Aida, M. Takeichi and N. Doi: "Effects of IP packet loss and picture frame reduction on MPEG1 subjective quality". IEEE 1999 Workshop on Multimedia Signal Processing, Copenhagen, 13-15 September 1999.

[63] IETF RFC 793: "Transmission Control Protocol".

[64] IETF RFC 2960: "Stream Control Transmission Protocol".

[65] IETF RFC 3286: "An Introduction to the Stream Control Transmission Protocol (SCTP)".

[66] T. V. Lakshman and U. Madhow: "The performance of TCP/IP for networks with high bandwidth-delay products and random loss". IEEE/ACM Transaction on Networking, vol. 5, No. 3, 1997, pp. 336-350.

[67] T. V. Lakshman, U. Madhow and B. Suter: "TCP/IP Performance with Random Loss and Bidirectional Congestion". IEEE/ACM Transaction on Networking, vol. 8, No. 5, October 2000, pp. 541-555.

[68] M. Mathis, J. Semke, J. Madhavi and T. Ott: "The Macroscopic Behaviour of the TCP Congestion Avoidance Algorithm". Computer Communication Review, vol 27, No. 3, 1997.

[69] J. Padhye, V. Firoiu, D. Towsley and J. Kurose: "Modeling TCP Throughput: A Simple Model and its Empirical validation". Proceedings of SIGCOMM'98.

[70] J. Nielsen: "Usability Engineering". Morgan Kauffman, San Francisco, 1994. ISBN 0-12-518406-9.

[71] B. G. C. Dellaert and B. E. Kahn: "How Tolerable is Delay? Consumers" Evaluation of Internet Web Sites after Waiting". Journal of Interactive Marketing vol 13, No. 1, 1999, pp 41-54.

[72] S. Khirman and P. Henriksen: "Relationship between Quality of Service and Quality of Experience for Public Internet Service". Passive and Active Measuring Conference (PAM 2002), March 2002.

[73] Draft ITU-T Recommendation G.1030 (2005): "Estimating end to end performance in IP networks for data applications".

[74] R. Steinmetz: "HumanPerception of Jitter and media Synchronization". IEEE Journal on Selected Areas in Communication vol. 14. no. 1 January 1996, pp. 61-72.

[75] ETSI ETR 297 (1996): "Human Factors (HF); Human Factors in Videotelephony".

[76] Q. Summerfield (1992): "Lipreading and audio-visual speech perception". Philosophical Transactions of the Royal Society of London, B335, pp. 71-78.

[77] P. Jardetzky, C. Sreenan and R. Needham (1995): "Storage and synchronisation for distributed continuous media. Multimedia Systems", 3, pp. 151-161.

[78] D. Lewkowicz (1996): "Perception of auditory-visual temporal synchrony in human infants". Journal of Experimental Psychology: Human Perception and Performance, 22(5), 1094-1106.

[79] ITU-R Recommendation BT.1359-1: "Relative timing of sound and vision for broadcasting".

[80] M. P. Hollier and R. Voelcker: "Towards a multi-modal perceptual model". BT Technology Journal vol 14, No 4., October 1997, pp. 162-171.

[81] M. P. Hollier, A. N. Rimell, D. S. Hands and R. Voelcker: "Multi-modal perception". BT Technology Journal vol. 17, No. 1, January 1999, pp. 35-46.

[82] J. G. Beerends and F. E. de Caluwe: "The Influence of Video Quality on Perceived Audio Quality and Vice Versa". Journal of Audio Engineering Society, vol. 47, No, 5, May 1999, pp 355-362.

[83] ITU-T Recommendation P.930 (1996): "Principles of a reference impairment system for video".

[84] D. S. Hands: "A basic Multimedia quality model". IEEE Transactions on multimedia, Vol. 6, No. 6, December 2004, pp. 806-816.

[85] S. Winkler and C. Faller: "Audiovsual Quality Evaluation of Low-Bitrate Video". IS&T/SPIE International Symposium on Electronic Imaging 2005, San Jose, January, 16-20, 2005.

[86] S. Winkler and C. Faller: "Maximizing Audiovisual Quality at Loe Bitrates". First International Workshop on Video Processing and Quality Metrics for Consumer Electronics Doubletree Paradise Valley Resort Scottsdale, Arizona, U.S.A., January, 23-25, 2005.

[87] N. Kitawaki, Y. Arayama and T. Yamada: "Multimedia opinion model based on media interaction of audio-visual communication". MESAQIN 2005 Measurement of Speech and Audio Quality in Networks, Prague, June, 9-110, 2005.

3 Definitions and abbreviations

3.1 Definitions

For the purposes of the present document, the following terms and definitions apply:

audio: all signals that are audible to human beings, including speech and music

codec: encoder/decoder pair

MultiMedia (MM): combination of two or more of the components speech, audio, video, data, with a temporal relationship between the various components

multimedia service application: service that handle several types of media such as audio and video in a synchronized way from the user's point of view

NOTE: A multimedia service application may involve multiple parties, multiple connections, and the addition or deletion of resources and users within a single communication session.

Quality of Experience (QoE): user perceived experience of what is being presented by a communication service or application user interface

NOTE: This definition is from TR 102 274 [1].

Quality of Service (QoS): the collective effect of service performance, which determine the degree of satisfaction of a user of the service

NOTE 1: The quality of service is characterized by the combined aspects of service support performance, service operability performance, serveability performance, service security performance and other factors specific to each service.

NOTE 2: The term "quality of service" is not used to express a degree of excellence in a comparative sense nor is it used in a quantitative sense for technical evaluations. In these cases a qualifying adjective (modifier) should be used.

NOTE 3: The definition above including notes 1 and 2 is from ITU-T Recommendation E.800 [2]. ITU-T Recommendation G.1000 [3] expands the definitions of QoS given in ITU-T Recommendation E.800 [2].

Round Trip Time (RTT): time required for a network communication to travel from the source to the destination and back

service component: part of a service, which describes a mono-medium communication related to a single information type

speech: oral production of information by a human being

streaming: mechanism whereby media content can be rendered at the same time that it is being transmitted to the client over the network

3.2 Abbreviations

For the purposes of the present document, the following abbreviations apply:

3GPP	3rd Generation Partnership Project
AAC	Advanced Audio Coding
ADPCM	Adaptive Differential Pulse Code Modulation
ANSI	American National Standards Institute
AVC	Advanced Video Coding
CELP	Code-Excited Linear Predictive
CIF	Common Intermediate Format
CN	Core Network
DMOS	Differential Mean Opinion Score
DPCM	Differential Pulse Code Modulation
DVD	Digital Versatile Disc
FEC	Forward Error Control
FR	Full Reference
GW	GateWay
IETF	Internet Engineering Task Force
iLBC	internet Low Bitrate Codec
IPDV	IP packet delay variation
IPER	IP packet error ratio
IPLR	IP packet loss ratio
IPTD	IP Packet Transfer Delay
ISDN	Integrated Services Digital Network
ITU-R	International Telecommunication Union - Radiocommunication Sector
ITU-T	International Telecommunication Union - Telecommunication Sector
LBR	Low Bit-rate Redundancy
LPC	Linear Predictive Coding
MM	MultiMedia
MOS	Mean Opinion Score
MP3	MPEG 1/2 Layer 3
MPEG	Moving Picture Experts Group
MT	Mobile Terminal
NAL	Network Adaptation Layer
NGN	Next Generation Network
ODG	Objective Difference Grade
PCM	Pulse Code Modulation
PEAQ	Perceptual Evaluation of Audio Quality
PLC	Packet Loss Concealment
QCIF	Quart CIF
QoE	Quality of Experience
QoS	Quality of Service
RACS	Resource and Admission Control Sub-System
RAN	Radio Access Network
RFC	Request For Comment
RTP	Real-Time Transport Protocol
RTT	Round Trip Time
SCTP	Stream Control Transmission Protocol
SIF	Standard Intermediate Format
SMPTE	Society of Motion Picture Television Engineers
TCP	Transport Control Protocol
TE	Terminal Equipment

TIA/EIA	Telecommunications Industry Association/Electronics Industry Association
UDP	User Datagram Protocol
UMTS	Universal Mobile Telecommunications System
UNI	User-Network Interface
VCL	Video Coding Layer
VQEG	Video Quality Experts Group

4 Framework for MM QoS classification

4.1 End user QoS requirements

Generally, end users care about the issues that are perceptible to them, only. The involvement of the user leads to the following conclusions from the end-user point of view:

- only the QoS perceived by end-user matter;
- quality criteria should be defined from the functional criteria, and then translated into technical criteria;
- the number of user defined/controlled attributes has to be as small as possible;
- derivation/definition of QoS attributes from the application requirements has to be simple;
- QoS attributes should be able to support all applications that are used, a certain number of applications have the characteristic of asymmetric nature between two directions, uplink/downlink;
- QoS definitions have to be future proof;
- QoS has to be provided end-to-end.

4.2 NGN QoS framework and requirements

TS 185 001 [4] provides a release framework for QoS in NGN and describes the requirements, which should be applied. The document addresses:

- QoS generic concepts.
- A QoS framework model.
- Various QoS classes regimes and mapping between them.
- Codec issues.
- QoS scenarios.
- QoS architecture requirements.
- QoS signalling requirements.

The functional architecture for the Resource and Admission Control Sub-System (RACS), in TISPAN NGN Release 1 is defined in ES 282 003 [5]. RACS is the TISPAN NGN subsystem responsible for elements of policing control including resource reservation and admission control. Two architectures for dynamic control of QoS are considered:

- Guaranteed QoS implying service delivery with absolute bound on some or all the QoS parameters.
- Relative QoS implying traffic class differentiation.

4.3 Classification made by other standards organizations

4.3.1 ITU-T

ITU-T Recommendations P.911 [6] addresses subjective audiovisual quality assessment methods for multimedia applications. In an annex to the Recommendation audio and video classes and attributes are described. These are summarized in table 1.

ITU-T Recommendation G.1010 [7] defines a model for multimedia Quality of Service (QoS) categories from an end-user viewpoint. By considering user expectations for a range of multimedia applications, eight distinct categories are identified, based on tolerance to information loss and delay. It is intended that these categories form the basis for defining realistic QoS classes for underlying transport networks, and associated QoS control mechanisms. Figure 1 describes this model.

In an informative annex to ITU-T Recommendation G.1010 [7] indications of suitable performance targets for audio, video and data applications are given. These targets are reproduced in tables 2 and 3.

Table 1: QoS classes for MM applications

QoS class	Description	
	Audio	**Video**
TV0	Studio quality, 20 bits/sample, 48 kHz loss-less, linear PCM coded audio.	Loss-less: ITU-R Recommendation BT.601 [8], 8 bits/sample linear PCM coded video in 4:2:2, Y, C_R, C_B format, video used for applications without compression.
TV1	Used for complete post-production, many edits and processing layers, intra-plant transmission. Also used for remote site plant transmission. Perceptually transparent when compared to TV0.	Used for complete post-production, many edits and processing layers, intra-plant transmission. Also used for remote site plant transmission. Perceptually transparent when compared to TV0.
TV2	Primary distribution: Used for simple modifications, few edits, programme mixing, and inter-facility transmission. Examples same as for video. Nearly perceptually transparent when compared to TV0.	Used for simple modifications, few edits, character/logo overlays, programme insertion, and inter-facility transmission. A broadcast example would be network-to-affiliate transmission. Other examples are a cable system regional downlink to a local head-end and high quality video conferencing system. Nearly perceptually transparent when compared to TV0.
TV3	Examples same as for video. Low audible artifacts are present when compared to TV2.	Used for delivery to home/consumer (no changes). Other examples are a cable system from the local head-end to a home and medium to high quality video conferencing. Low artifacts are present when compared to TV2.
MM4	Low audible artifacts relative to TV3 using speech and audio. Medium quality video conferencing. Usually full audio bandwidth (20 Hz to 20 000 Hz), but bandwidths down to wideband (50 Hz to 7 000 Hz) are acceptable.	All frames encoded. Low artifacts relative to TV3. Medium quality video conferencing. Usually ≥ 25 fps for 625-lines systems and ≥ 30 fps for 525 lines systems.
MM5	Low audible artifacts relative to a narrow-band reference (300 Hz to 3 400 Hz telephony band) using speech and music. Perceptual artifacts possible, but quality level useful for designed tasks, e.g. low quality video conferencing.	Frames may be dropped at encoder. Perceivable artifacts possible, but quality level useful for designed tasks, e.g. low quality video conferencing.
MM6	Severe audible artifacts relative to a narrow-band (300 Hz to 3 400 Hz) telephony application. Speech is however still intelligible.	Series of stills. Not intended to provide full motion (examples: surveillance, graphics).

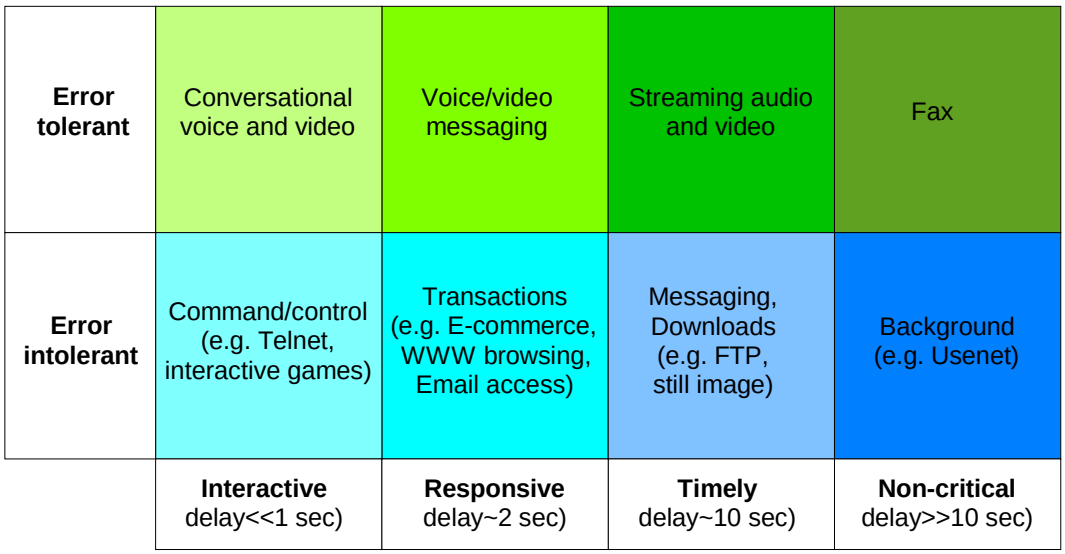


Figure 1: ITU-T Recommendation G.1010 [7] model for user-centric QoS categories

Table 2: Performance targets for audio and video applications, ITU-T Recommendation G.1010 [7]

Medium	Service Application	Degree of symmetry	Typical data rates	Key performance parameters and target values			
				One-way delay	Delay variation	Information loss (note 2)	Other
Audio	Conversational voice e.g. telephony	Two-way	4 kb/s to 64 kb/s	< 150 msec preferred (note 1) < 400 msec limit (note 1)	< 1 msec	< 3 % packet loss ratio (PLR)	
Audio	Voice messaging	Primarily one-way	4 kb/s to 32 kb/s	< 1 sec for playback < 2 sec for record	< 1 msec	< 3 % PLR	
Audio	High quality streaming audio	Primarily one-way	16 kb/s to 128 kb/s (note 3)	< 10 sec	<< 1 msec	< 1 % PLR	
Video	Videophone	Two-way	16 kb/s to 384 kb/s	< 150 msec preferred (note 4) <400 msec limit		< 1 % PLR	Lip-synch : < 80 msec
Video	Broadcast	One-way	16 kb/s to 384 kb/s	< 10 sec		< 1 % PLR	

NOTE 1: Assumes adequate echo control.
NOTE 2: Exact values depend on specific codec, but assumes use of a packet loss concealment algorithm to minimize effect of packet loss.
NOTE 3: Quality is very dependent on codec type and bit-rate.
NOTE 4: These values are to be considered as long-term target values which may not be met by current technology.

Table 3: Performance targets for data applications, ITU-T Recommendation G.1010 [7]

Medium	Service Application	Degree of symmetry	Typical amount of data	Key performance parameters and target values		
				One-way Delay	Delay variation	Information loss
Data	Web-browsing - HTML	Primarily one-way	~10 kB	Preferred < 2 sec /page Acceptable < 4 sec/page	N.A	Zero
Data	Bulk data transfer/retrieval	Primarily one-way	10 kB to 10 MB	Preferred < 15 sec Acceptable < 60 sec	N.A	Zero
Data	Transaction services – high priority e.g. e-commerce, ATM	Two-way	< 10 kB	Preferred < 2 sec Acceptable < 4 sec	N.A	Zero
Data	Command/control	Two-way	~ 1 kB	< 250 msec	N.A	Zero
Data	Still image	One-way	< 100 kB	Preferred < 15 sec Acceptable < 60 sec	N.A	Zero
Data	Interactive games	Two-way	< 1 kB	< 200 msec	N.A	Zero
Data	Telnet	Two-way (asymmetric)	< 1 kB	< 200 msec	N.A	Zero
Data	E-mail (server access)	Primarily One-way	< 10 kB	Preferred < 2 sec Acceptable < 4 sec	N.A	Zero
Data	E-mail (server to server transfer)	Primarily one-way	< 10 kB	Can be several minutes	N.A	Zero
Data	Fax ('real-time')	Primarily one-way	~ 10 kB	< 30 sec/page	N.A	< 10-6 BER
Data	Fax (store and forward)	Primarily one-way	~ 10kB	Can be several minutes	N.A	< 10-6 BER
Data	Low priority transactions	Primarily one-way	< 10 kB	< 30 sec	N.A	Zero
Data	Usenet	Primarily one-way	Can be 1 MB or more	Can be several minutes	N.A	Zero
NOTE: In some cases, it may be more appropriate to consider these values as response times.						

ITU-T Recommendation Y.1541 [9] specifies network performance objectives for IP based services. The scope of the Recommendation is the UNI to UNI Interfaces as described in figure 2, user equipments characteristics are not included.

Table 4 presents an overview network QoS class definitions and network performance objectives.

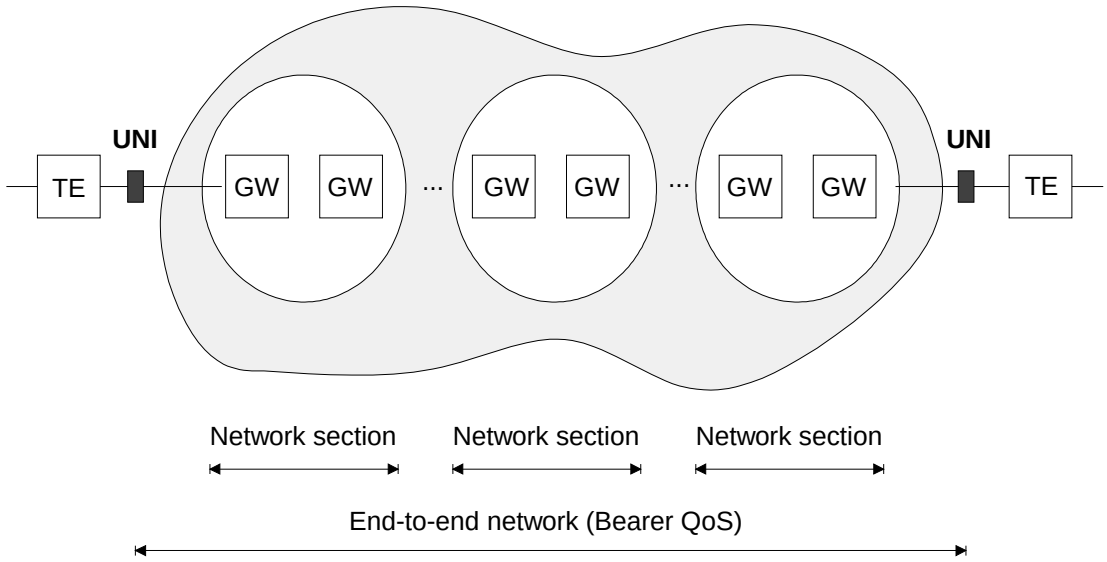


Figure 2: Reference path for network QoS objectives defined in ITU-T Recommendation Y.1541 [9]

NOTE 1: Versions of ITU-T Recommendation Y.1541 [9] published after May 2002 may define more classes.

Table 4: IP network QoS class definitions and network performance objectives ITU-T Recommendation Y.1541 [9]

		QoS Classes					
Network Performance Parameter	**Nature of Network Performance Objective**	**Class 0**	**Class 1**	**Class 2**	**Class 3**	**Class 4**	**Class 5 Un-specified**
IPTD	Upper bound on the mean IPTD (note 1)	100 msec	400 msec	100 msec	400 msec	1 sec	U
IPDV	Upper bound on the $1 - 10^{-3}$ quantile of IPTD minus the minimum IPTD (note 2)	50 msec (note 3)	50 msec (note 3)	U	U	U	U
IPLR	Upper bound on the packet loss probability	1×10^{-3} (note 4)	1×10^{-3} (note 4)	1×10^{-3}	1×10^{-3}	1×10^{-3}	U
IPER	Upper bound	1×10^{-4} (note 5)					U

General Notes:
The objectives apply to public IP Networks. The objectives are believed to be achievable on common IP network implementations. The network providers' commitment to the user is to attempt to deliver packets in a way that achieves each of the applicable objectives. The vast majority of IP paths advertising conformance with ITU-T Recommendation Y.1541 [9] should meet those objectives. For some parameters, performance on shorter and/or less complex paths may be significantly better.
An evaluation interval of 1 minute is provisionally suggested for IPTD, IPDV, and IPLR, and in all cases the interval must be reported.
Individual network providers may choose to offer performance commitments better than these objectives.
"U" means "unspecified" or "unbounded". When the performance relative to a particular parameter is identified as being "U" the ITU-T establishes no objective for this parameter and any default Y.1541 objective can be ignored. When the objective for a parameter is set to "U", performance with respect to that parameter may, at times, be arbitrarily poor.
All values are provisional and they need not be met by networks until they are revised (up or down) based on real operational experience.

NOTE 1: Very long propagation times will prevent low end-to-end delay objectives from being met. In these and some other circumstances, the IPTD objectives in Classes 0 and 2 will not always be achievable. Every network provider will encounter these circumstances and the range of IPTD objectives in table 1 provides achievable QoS classes as alternatives. The delay objectives of a class do not preclude a network provider from offering services with shorter delay commitments. According to the definition of IPTD in ITU-T Recommendation Y.1540 [10], packet insertion time is included in the IPTD objective. This Recommendation suggests a maximum packet information field of 1 500 bytes for evaluating these objectives.

NOTE 2: The definition and nature of the IPDV objective is under study. See appendix II of ITU-T Recommendation Y.1541 [9] for more details.

NOTE 3: This value is dependent on the capacity of inter-network links. Smaller variations are possible when all capacities are higher than primary rate (T1 or E1), or when competing packet information fields are smaller than 1 500 bytes (see appendix IV ITU-T Recommendation Y.1541 [9]).

NOTE 4: The Class 0 and 1 objectives for IPLR are partly based on studies showing that high quality voice applications and voice codecs will be essentially unaffected by a 10.3 IPLR.

NOTE 5: This value ensures that packet loss is the dominant source of defects presented to upper layers, and is feasible with IP transport on ATM.

NOTE 2: The appendixes referred to in table 4 are all appendixes to ITU-T Recommendation Y.1541 [9].

NOTE 3: Versions of ITU-T Recommendation Y.1541 [9] published after May 2002 may define more classes.

ITU-T Recommendation G.1050 [12] describes a model for evaluating transmission performance over an IP network. The model is originally developed by TIA TR-30. The Recommendation focuses on the impact of impairments on Layer 3 performance. IP streams from any type of network can be evaluated using the model.

It is stated that the model will continue to evolve as more information becomes available.

The following examples of types of equipment that can evaluated are listed:

- IP-connected endpoints:
  - IP Network Devices (such as: User Agents, Call Agents, Media Servers, Media Gateway Controllers, Gatekeepers, Application Servers, Edge Routers, etc.).
  - IP Video.

  - IP Phones.
  - IAF (Internet Aware Fax).
- PSTN-connected devices through IP gateways:
  - POTS through Voice-over-IP (VoIP) gateways.
  - T.38 facsimile devices and gateways.
  - V.150.1 and V.152 (voiceband data, VBD) modem-over-IP gateways.
  - V.151 textphone-over-IP gateways.

The Recommendation indicates the following limitations:

- The IP Network Model is not intended to represent any specific IP Network. Rather it provides a range of test scenarios that could represent a wide range of IP network characteristics, such as those experienced in Well-Managed (QoS managed networks), Partially-Managed (Non-QoS) and Un-managed (Internet) networks.
- The model only addresses the IP portion of the network and does not address the PSTN portion of a connection that may utilize PSTN at one or both ends of the connection.
- The Recommendation does not model all possible connections that can be encountered between devices.
- The IP network model is based on an informal survey of anonymous IP service providers and IP network equipment manufacturers in the 2005 timeframe and will continue to evolve as more statistical information becomes available and as the IP network evolves.

ITU-T Recommendation H.360 [11] contains reference architecture for controlling the QoS and service priority of multimedia services in networks, which are comprised of combinations of switched circuit and packet domains, wireless and wireline technologies, and conventional and packet-based terminals. The reference architecture is functionally defined. The Recommendation uses a domain-based approach, which allows issues of administrative control and security also to be considered. This Recommendation is concerned with end-to-end QoS control in multimedia systems made up of multiple and disparate administrative domains and QoS mechanisms.

ITU-T Recommendation SG 16 is also working on a Recommendation that provides the enhancements to the H.323 signalling to allow a call to establish end-to-end QoS and service priority.

4.3.2 3GPP

In TS 123 107 [13] the UMTS QoS classes, also referred to as traffic classes, are defined. The document specifies the list of attributes applicable to UMTS Bearer Service and Radio Access Bearer Service, as well as describes the Quality of Service architecture to be used in UMTS networks. The UMTS QoS Architecture is described in figure 4. The scope of TS 123 107 [13] is restricted to the UMTS bearer service as illustrated in figure 3, i.e. between the Mobile Termination (MT) and the Core Network (CN) Gateway.

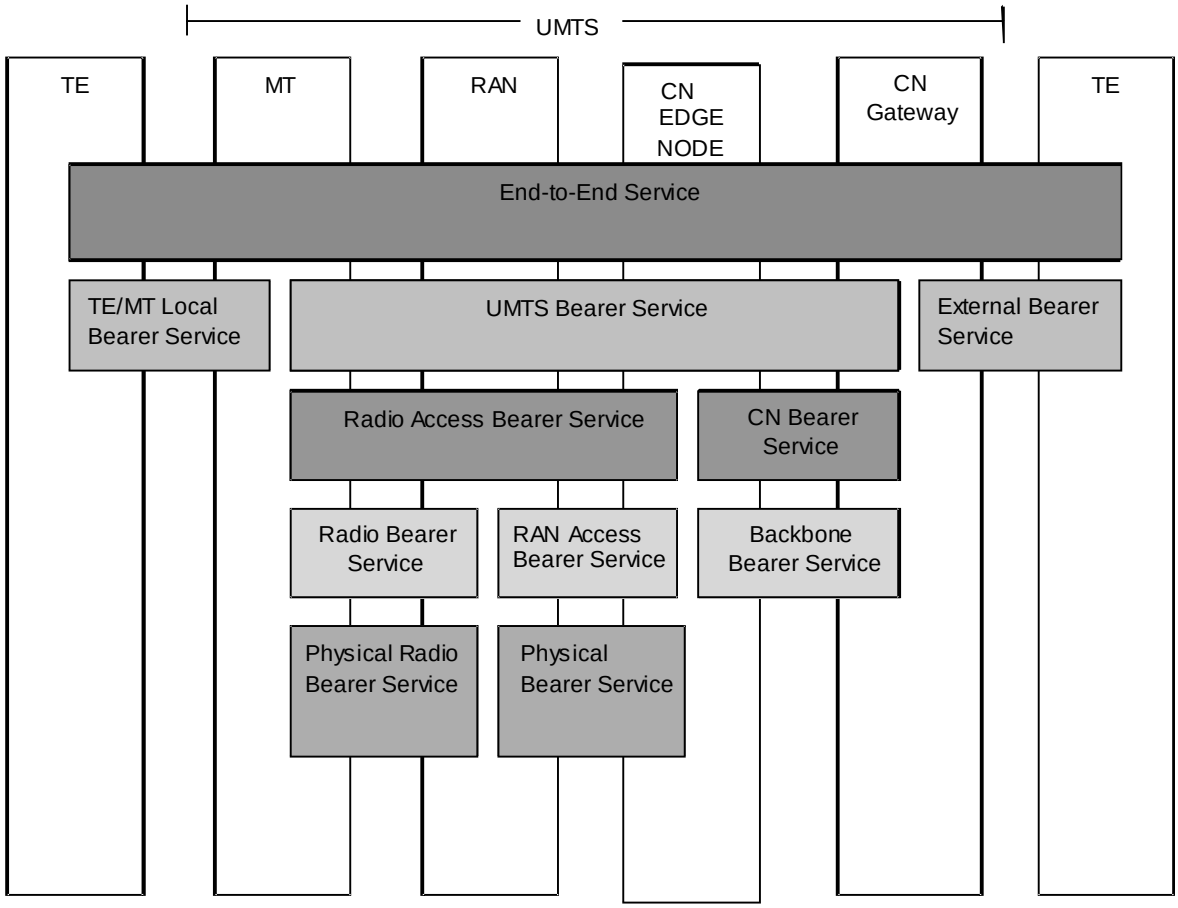


Figure 3: UMTS QoS architecture

There are defined four different QoS classes:

- conversational class;
- streaming class;
- interactive class;
- background class.

Table 5 illustrates these QoS classes.

Table 5: UMTS QoS classes

Traffic class	Conversational class conversational RT	Streaming class streaming RT	Interactive class Interactive best effort	Background Background best effort
Fundamental characteristics	- Preserve time relation (variation) between information entities of the stream Conversational pattern (stringent and low delay)	- Preserve time relation (variation) between information entities of the stream	- Request response pattern - Preserve payload content	- Destination is not expecting the data within a certain time - Preserve payload content
Example of the application	- Voice	- Streaming video	- Web browsing	- Background download of emails

The UMTS bearer attributes and their relevancy for each bearer traffic class are summarized in table 6.

Table 6: UMTS bearer attributes defined for each bearer traffic class

Traffic class	Conversational class	Streaming class	Interactive class	Background class
Maximum bitrate	X	X	X	X
Delivery order	X	X	X	X
Maximum SDU size	X	X	X	X
SDU format information	X	X		
SDU error ratio	X	X	X	X
Residual bit error ratio	X	X	X	X
Delivery of erroneous SDUs	X	X	X	X
Transfer delay	X	X		
Guaranteed bit rate	X	X		
Traffic handling priority			X	
Allocation/Retention priority	X	X	X	X
Source statistics descriptor	X	X		
Signalling indication			X	

4.3.3 ANSI

ANSI T1.522 [14] specifies the classes of QoS sufficient to support Business Multimedia Conferencing on IP networks, defined as equivalent to legacy conference system performance (e.g. ITU-T Recommendation H.320 [49] systems). In the present document two baseline conferencing systems are defined:

- Tier 1 Desktop PC Systems.
- Tier 2 Group Conference Room Systems.

The specification defines perceptible performance level and acceptable performance level. For some aspects, e.g. bit rate and loss, there are different levels specified for Tier 1 and for Tier 2.

The performance parameters of interest are summarized in table 7 and table 8. The parameters specified in ANSI T1.522 [14] are:

- Delay (Both at User interface and Network interface).
- Delay variation.
- Media synchronization.
- Accessibility.
- Lost Transport Packet Rate.
- Dropped connection.
- Committed bit rate.

Table 7: Quality of Service Parameters - MM user Interface

	Quality Criteria		
Communication Function	**Speed**	**Accuracy**	**Availability and Reliability**
Connection Establishment	Set-up time Transfer time	Mis-directed	Accessibility ratio (among media) Connection Failure
User information Transfer	Delay (Spontaneity) Delay variation Contention Resolution	Media Quality Media Synchronization	Dropped Connection
Connection release	Take down time		Release failure

Table 8: Quality of Service Parameters - End-to-end Interface

	Quality Criteria		
Communication Function	**Speed**	**Accuracy**	**Availability and Reliability**
Connection Establishment	Set-up time Transfer time	Mis-directed	Accessibility Connection Failure
User information (Packet) Transfer	Delay (Network latency) Delay variation (within a single media stream and between streams) Information Bit rate (Committed Bit rate and Delivered Bit Rate)	Lost Transport Packet Rate (combines IP Packet Defects, such as Errored and Lost Packets)	Dropped Connection (IP Availability)
Connection release	Take-down time		Release failure

5 Multimedia QoS performance metrics

5.1 End-to-end performance characteristics for speech service component

5.1.1 Speech coding algorithms

A list of standardized speech and video codecs for conversational applications is presented in annex 1 to TS 185 001 [4]. Most of these codecs are based on an 8 kHz sampling rate, which means that the maximum speech bandwidth that can be transmitted by these codecs is 4 kHz. Telephone bandwidth is usually somewhat less, 300 Hz to 3 400 Hz. Speech transmitted on such a channel is often characterized as narrowband speech.

There is also a group of speech coding algorithms that is based on 16 kHz sampling enabling transmission of speech signals covering the frequency band from 50 Hz to 7 000 Hz. These codecs are often characterized as wideband codecs, and are using a 16 bit A/D converter. The dynamic range of these codecs is therefore larger than the dynamic range of the narrowband codecs.

Speech codecs are often divided into three classes:

- waveform codecs;
- source codecs;
- hybrid codecs.

Waveform codecs attempt, without using any knowledge of how the signal to be coded was generated, to produce a reconstructed signal whose waveform is as close as possible to the original. This means that in theory they should be signal independent and work well with non-speech signals. An example is the codec standardized in ITU-T Recommendation G.711 [15].

To reduce the required bitrate, the difference compared with the previous sample may be transmitted instead of the actual sample. This technique is call delta modulation or differential PCM (DPCM). This technique may be further enhanced by predicting the value of the next sample from the previous samples and transmit the difference between the predicted value and the actual sampled value (ADPCM).

The input speech signal may also be split into a number of frequency bands, or sub-bands, and each is coded independently. This is called Sub-band coding. An example is the codec defined in ITU-T Recommendation G.722 [16] where the 7 kHz frequency band is divided into two sub-bands, which are coded independent of each other.

Source coders operate using a model of how the source was generated, and attempt to extract, from the signal being coded, the parameters of the model. Coders using this technique require very low bitrate, but the quality is not good enough for public telecommunication applications.

Hybrid codecs attempt to fill the gap between waveform and source codecs. Although other forms of hybrid codecs exist, the most successful and commonly used are time domain Analysis-by-Synthesis (AbS) codecs. Such coders use the same linear prediction filter model of the vocal tract as found in LPC vocoders. However instead of applying a simple two-state, voiced/unvoiced, model to find the necessary input to this filter, the excitation signal is chosen by attempting to match the reconstructed speech waveform as closely as possible to the original speech waveform. Examples are Multi-Pulse Excited (MPE) codecs and Code-Excited Linear Predictive (CELP) codecs.

There are speech codecs that provides embedded voice processing solutions for real-time communications on packet networks. One of these, iLBC, is the subject of RFC 3951 [17]. A paper by Andersen et al. [18] presents an overview of the algorithm. The algorithm is more robust to packet loss than the codecs standardized by ITU-T and 3GPP/ETSI.

ITU-T has recently begun work on a wideband extension to ITU-T Recommendation G.729 [19] allowing interworking between the narrowband and wideband modes.

5.1.2 Delay and jitter

The delay sources of a multimedia connection on an IP network are:

- Transmitting terminal delay.
  The main sources are the signal processing (Codec) and the signal packetization.
- Network delay.
- Receiving terminal delay.
  The main source is the receive jitter buffer.

The speech coding related delay depends on whether the coding algorithm is sample-based or frame-based.

The **sample-based algorithms** are low-delay algorithms, introducing less than 10 msec delay (usually 3 msec or less). The most common sample-based algorithm in telecommunications is PCM defined in ITU-T Recommendation G.711 [15].

The **frame-based algorithms** segment the speech signals into frames that typically are 20 msec long. These frames are processed using various techniques (e.g. CELP or MPE). Although 20 msec is the most commonly used frame size, it should be noted there are standardized algorithms that use 10 msec or 30 msec frames.

To reduce the effect of packet loss, Forward Error Control/Packet Loss Concealment can be used. To do so some codecs include an extra time window called look-ahead.

The delay introduced by a **frame-based algorithm** is:

$$2 \times \text{frame size} + \text{look-ahead} \quad (1)$$

The duration of a voice packet is rather flexible. RFC 3551 [20], the IETF Standard that defines the profiles for the Real Time Protocol (RTP) defined in RFC 3550 [21], recommends a packet duration of 20 msec except for ITU-T Recommendation G.723.1 [22]. For ITU-T Recommendation G.723.1 30 msec is recommended. This is because the packet duration has to be a multiple of the coding algorithm frame size.

For the sample-based algorithms the delay introduced by an application is equal to the packet duration plus an algorithm delay of 3 msec to 10 msec, depending on the algorithm.

If multiple voice frames belonging to a frame-based algorithm are grouped together into a single packet, the extra delay will be the duration of one voice frame for each additional voice frame added to the packet:

$$(N+1) \times \text{frame size} + \text{look-ahead} \quad (2)$$

where N is the number of voice frames in each packet.

The core network delay sources are the delay caused at each router of the network connection and the propagation delay.

TIA/EIA-TSB-116 [23] indicates that the router related delay is approx. 1,5 msec per hop.

The propagation delay depends on the technology used. Table A.1/G.114 of ITU-T Recommendation G.114 [24] presents planning values for calculating propagation delay for various transmission technologies.

The access network may be a significant delay contributor, subject to the technology used. As an example the delay introduced by ADSL may be more than 10 msec depending on the ADSL link capacity. For asymmetrical systems the delay upstream is larger than the delay downstream. For other access technologies it is likely that the delay is less.

As stated in the introduction of this clause, the main delay source at the receiving terminal is the jitter buffer. Jitter is defined as the delay variation caused by queuing in network elements or by routing the packets along different network paths. These delay variations need to be removed before replaying the audible signal to the human user. This is achieved by inserting a buffer (jitter buffer or playout buffer) at the receiving terminal.

The size of the jitter buffer needs to match the amount of jitter at the receiving terminal. When the jitter buffer is to short, packet may arrive too late and will be lost. On the other hand, long jitter buffer increases the end-to-end delay perceived by the user.

The jitter buffer size can be fixed or adaptive. In most scenarios an adaptive jitter buffer is preferable because the jitter characteristics may depend on the actual connection and traffic scenario.

Most implementations adjust the playout buffer size between talkspurts. However, experiments carried out by Liang et al. [25] show that is it possible to adjust the playout of each individual packet by scaling the packets. The packets can be scaled from 50 % to 200 % of their original size without degrading sound quality. This enables implementations to reduce the average delay compared with traditional techniques.

The effects of delay are described in ITU-T Recommendation G.114 [24]. The present version of the Recommendation presents E-model [26] calculation results where delay is one of the input parameters (see note). The Recommendation states that it is desirable to keep the delays seen by user applications as low as possible. Although a few applications may be slightly affected by end-to-end delays of less than 150 msec, if delays can be kept below this figure, most applications, both speech and non-speech, will experience essentially transparent interactivity. The upper delay limit for planning purposes is 400 msec. It is however recognized that in some exceptional cases (e.g. double satellite hops) this limit will be exceeded.

NOTE: A description of the E-model is given in clause 5.1.4.

Annex B to the year 2000 version of ITU-T Recommendation G.114 [24] presents an overview of test results addressing the effects of delay on interactive voice communication.

5.1.3 Packet loss

Packet loss degrades the perceived speech quality. The amount of degradation depends on the robustness of the speech codec, and whether or not protection mechanisms such as Packet Loss Concealment (PLC) are implemented. Test results presented to TS 101 329-6 [27] illustrate both the degradation caused by packet loss and the effects of PLC. Figure 4 describes the effects of packet loss on MOS (Mean Opinion Score) for some relevant codecs with and without PLC.

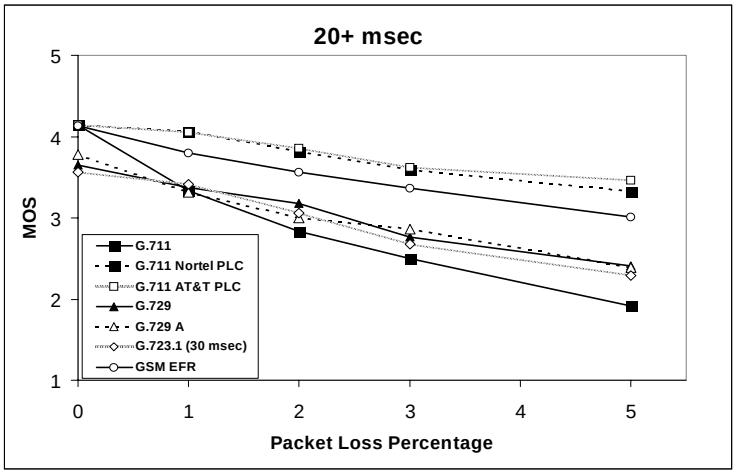


Figure 4: Effects of packet loss on voice quality (MOS) [27]

The tests described above are made when single packets are lost with random distribution. On real networks there are quite frequently bursts of packets lost, not single packets, due to effects such as network overload, router queuing and radio transmission disturbances. Subject to the size of the burst (number of consecutive packets lost), burst packet loss may degrade the voice quality more than single packet loss. A paper by Clark [28] presents a review of the effect of burst packet loss. The effect is illustrated in figure 5. The codec used and the burst generating method are not identified in the paper.

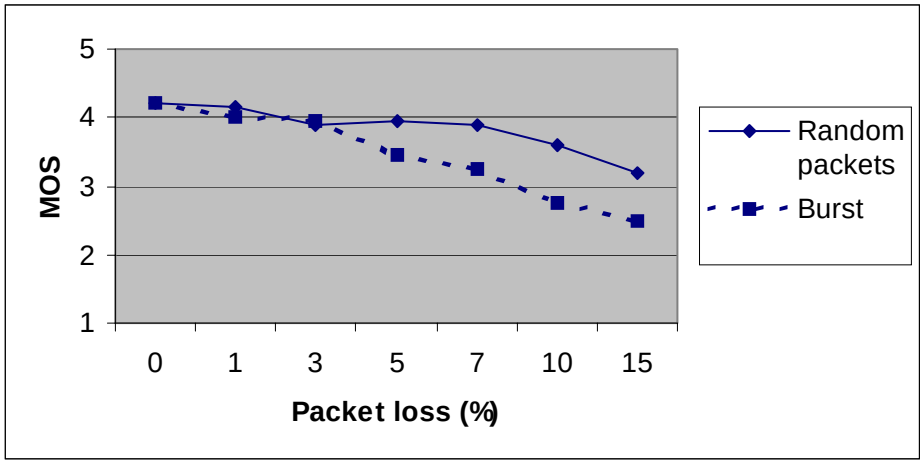


Figure 5: Effects of burst packet loss on MOS [28]

A paper by Jiang and Schulzrinne [29] compares packet loss repair methods and effect on perceived voice quality under bursty loss. Both calculation results using the E-model [26] and results from subjective tests are reported.

To generate burst packet loss the Gilbert model is used. There are indications that this model is too simple, in TS 101 329-5 [30] a four-state Markov model is suggested. (The Gilbert model is a two-state model).
The relation between the (average) loss probability, the conditional loss probability (p_c) and the unconditional loss probability (p_u) is,

$$p = \frac{p_u(1-p_c)}{1-p_u} \tag{3}$$

The conditional loss probability describes the loss burstiness.

Most of the tests were carried out using the ITU-T Recommendation G.729 [19] speech coding algorithm.

The basic loss pattern is specified at 20 msec packet interval. To correctly simulate the loss pattern at larger packet intervals, every second event is picked to simulate 40 msec packet intervals, every third event to simulated 60 msec packet intervals and so on. The consequence of this approach is decreased burstiness when the packet intervals increase.

Two packet loss repair methods were compared:

- Forward Error Correction (FEC).
- Low Bit-rate Redundancy (LBR) where a redundant but lover quality version of the same signal is transmitted.

When no packet loss repair mechanism is implemented the test results show that the difference between the random loss rating and the bursty loss rating is between 0,2 and 0,4 on a five-point MOS rating scale as illustrated in figure 6.

Another observation made is that 40 msec packet interval is rated better than 20 msec packet interval. However, the tests described are listening test, not including delay effects. The paper [29] presents an analysis of delay effects using the E-model [26], and concludes that 20 msec and 40 msec packet intervals could be considered as equal, while the performance when using longer packet intervals is worse.

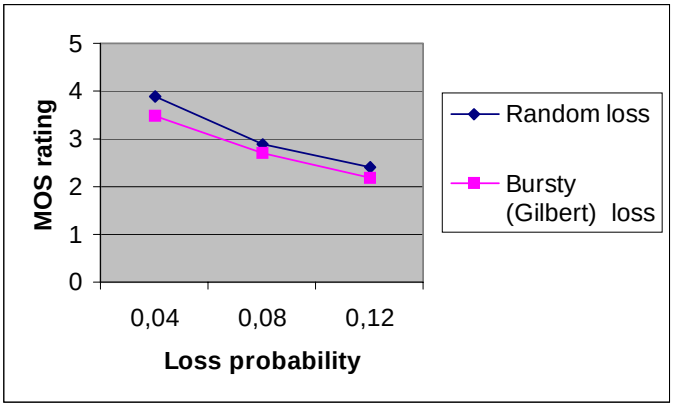


Figure 6: MOS rating, random loss and bursty (Gilbert) loss, Packet interval 30 msec, p_c= 0,3 [29]

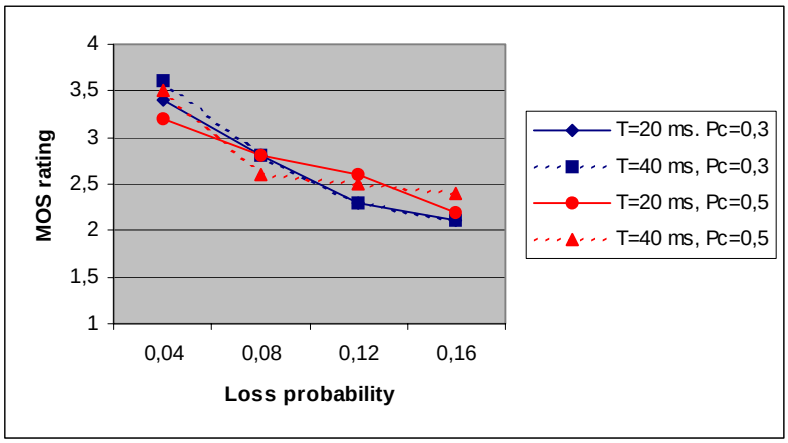


Figure 7: MOS rating, bursty (Gilbert) loss 20 msec and 40 msec packet intervals [29]

Jiang and Schulzrinne [29] conclude that FEC is better than LBR.

A presentation by Duric [31] describes the packet loss robustness of the iLBC codec. There are other non-standardized codecs having similar robustness, and robustness enhancement units that provide increased packet loss robustness for both standardized and non-standardized speech codecs.

5.1.4 Overall rating

A useful tool for estimating the effects of transmission degradation on voice quality is the E-model standardized in ITU-T Recommendation G.107 [26]. The E-model is based on work carried out by ETSI and documented in ETR 250 [33]. The results were presented to ITU-T. The model was accepted and has been improved by ITU-T Recommendation SG 12. TR 102 356 [34] presents an overview of most of these improvements, however this is an ongoing activity.

The model calculates a rating factor:

$$R = R_0 - I_s - I_d - I_{e-eff} + A \quad (4)$$

where

R_0 is related to the signal-to-noise ratio,
I_s is related to impairments such as loudness ratings and qdu,
I_d is related to delay,
$I_{e\text{-}eff}$ is related to impairments caused by low bitrate codecs,
A is an advantage factor.

For the time being the E-model is restricted to narrowband (3,1 kHz) voice.

5.2 End-to-end performance characteristics for audio service component

5.2.1 Audio coding algorithms

Standardized audio coding algorithms have been developed by the ISO/IEC Moving Picture Experts Group (MPEG), a group working on coding standards for audio and video. The group has produced three sets of standards; MPEG 1 [35], MPEG 2 [36] and MPEG 4 [37].

The first audio coding generation is defined in part 3 of MPEG 1 [35]. Three operating modes, called layers with increasing complexity and performance, are defined. Layer 3 is the highest complexity mode, optimized to provide the highest quality at bit-rates around 128 kbit/s for stereo signals. This coding algorithm is often referred to as MP3. MPEG 1 [35] Part 3 defines three sampling frequencies; 32 kHz, 44,1 kHz and 48 kHz.

MPEG 2 [36] extends the sampling frequencies to 16 kHz, 22,05 kHz and 24 kHz allowing for lower transmission bitrates. There is no modification to the algorithm.

A second generation audio coding, called Advanced Audio Coding (AAC) is defined in MPEG 2 [36] part 7. This is a completely new algorithm that is not backward compatible with MP3. AAC supports a wide range of sampling rates (8 kHz to 96 kHz). The range of bit rates supported is from 16 kbit/s up to 576 kbit/s and up to can be supported 48 audio channels.

MPEG 4 [37] extends the AAC algorithm with a low delay option. The algorithmic delay of this coding algorithm is 20 msec [37].

There is also a High Efficiency AAC (HE-AAC) profile [39]. HE AAC can deliver coding gains of more than 40 % compared to MPEG-4 AAC.

A paper by Brandenburg [40] presents an overview of the MP3 and AAC algorithms.

There are also some proprietary algorithms that are widely used, e.g.:

- Real Audio of Real Networks.
- Windows Media of Microsoft.

Both algorithms are used in streaming applications where the audio is stored on a server. Most of the information stored is multimedia; the audio is usually transmitted with video.

The latest version of ITU-T Recommendation G.722.1 [32] is specifying a 14 kHz bandwidth option.

5.2.2 Delay and jitter

To ensure that the audio data is played smoothly, without pauses between data fragments, most audio streaming applications use buffers to cope with both jitter and pauses in the media streams. Streaming has not the same interactivity requirements as telephony, and is not delay sensitive. The size of the buffer is therefore not critical, and has minimal effect on the user perceived quality.

The delay of the AAC algorithm is described in figure 8. While the delay of the basic AAC algorithm is depending on the bitrate (i.e. the signal compression), the Low Delay option has a fixed delay of 20 msec.

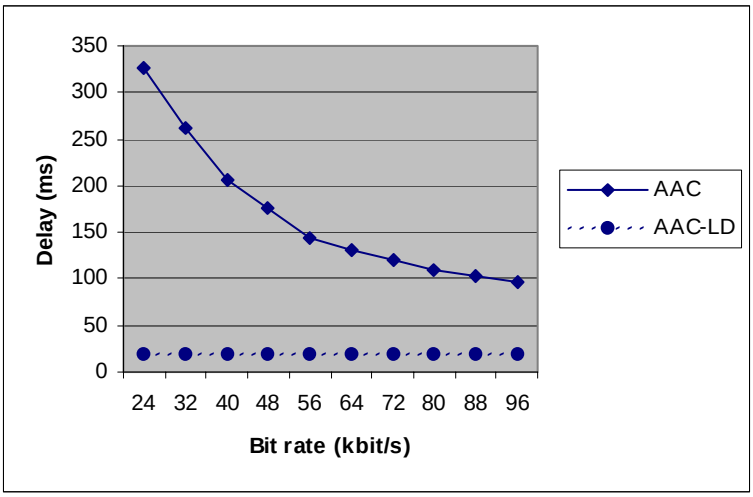


Figure 8: MPEG AAC algorithm delay [38]

5.2.3 Packet loss

Audio streaming may use TCP or UDP as the transport protocol. Where TCP is used, retransmission takes care of packet loss.

There is no information about the effects of packet loss on the audio codecs performance other than statements that the use of packet loss repair algorithms improves the performance.

5.2.4 Bit rate

While most of the speech coding algorithms discussed in clause 5.1 are transmitting at a single fixed bit rate, both the standardized and proprietary audio coding algorithms identified in clause 5.2.1 offer a great deal of flexibility to users and user applications. The lowest bitrate indicated is 4 kbit/s, however this bitrate is assuming 8 kHz sampling and is suitable for speech only. More common bitrates are in the range between 64 kbit/s and 192 kbit/s.

5.2.5 Overall rating

ITU-R Recommendation BS.1387 [41] describes a method for measuring the quality of wide bandwidth audio, Perceptual Evaluation of Audio Quality (PEAQ). The psychoacoustics model employed in the method produces a number of variables based on comparisons between a reference signal and the same signal processed by a particular device such as a codec. These variables are used to predict the subjective quality rating that would be assigned to the processed signal if a formal listening test were conducted. The algorithm calculates an Objective Difference Grade (ODG) that corresponds to the results of subjective tests where the listeners are asked to assess the impairments. The scale used is described in table 9.

Table 9: The ITU-R five grade impairment scale

Rating	Description
5	Imperceptible
4	Perceptible but not annoying
3	Slightly annoying
2	Annoying
1	Very annoying

5.3 End-to-end performance characteristics for video service components

5.3.1 Video coding algorithms

The video coding algorithms standardized in the ITU-T Recommendations H.261 [42], H.263 [43] and H.264 [44] are designed for interactive real-time communication (videotelephony and videoconferencing).
These standards are not explicitly defining a CODEC. Rather, they define the syntax of an encoded video bitstream together with the method of decoding this bitstream. The consequence is that there might significant quality differences between codecs conforming to the same standard.

ITU-T Recommendation H.261 [42] was a major step forward for videophone and multimedia applications offered on public switched networks (ISDNs). It was designed for data rates which are multiples of 64 Kbit/s, and is sometimes called p x 64 Kbit/s (p is in the range 1 to 30). There are however ITU-T Recommendation H.261 [42] based videophone implementations where audio and video are sent on a single 64 kbit/s channel. The video rate for these implementations is approx. 46 kbit/s. Implementation of the ITU-T Recommendation H.261 [42] codec is mandatory for end systems that claim conformance to ITU-T Recommendations H.320 [49] (ISDN) and H.323 [50] (Packet networks).

ITU-T Recommendation H.263 [43] was particularly designed for video coding at low bitrates (20 kbit/s or less). The ITU-T Recommendation H.263 [43] algorithm is basically based on the same principles as the ITU-T Recommendation H.261 [42] algorithm, but may give a better picture quality than ITU-T Recommendation H.261 [42] at the same bitrate. Furthermore, ITU-T Recommendation H.263 [43] supports a wider range of bitrates than ITU-T Recommendation H.261 [42], from approx. 9,6 kbit/s to 2 Mbit/s.

A second version of the standard, often called H.263+, added numerous coding options to improve codec performance over the first version. Côté et al. [53] presents an overview of the ITU-T Recommendation H.263 + video coding algorithm.

The latest video coding Recommendation is ITU-T Recommendation H.264 [44]. The main goals of the H.264/AVC standardization effort have been enhanced compression performance and provision of a "network-friendly" video representation addressing "conversational" (video telephony) and 'nonconversational' (storage, broadcast, or streaming) applications. A paper by Wiegand et al. [53] presents an overview of the ITU-T Recommendation H.264 [44] standard.

ITU-T Recommendation H.264 [44] consists of a Video Coding Layer (VCL) and a Network Adaptation Layer (NAL). This approach makes it easier to adapt the video coding stream to various network and multiplex environments. The transport of ITU-T Recommendation H.264 [44] coded video over IP networks is analysed in a paper by Wenger [53].

There are three profiles defined for H.264: Baseline, Main and Extended. The Baseline profile is nearly perfectly designed for video conferencing applications as it provides robust error resilience tools (which provide for good video quality even on error-prone networks such as the Internet) and allows for low latency coding and decoding, which makes video conferences feel more natural The Main and Extended profiles are better suited for television applications (digital broadcast, DVD), and for video streaming.

MPEG 1 [35] was initially designed for storage and transmission of media at up to 1,5 Mbit/s. Part 2 of the standard defines the video coding principles. The quality of the MPEG 1 [35] video coding is similar to the quality of a VHS recorder.

MPEG 2 [36] allows for coding of studio quality video for digital TV including High Definition TV. Like MPEG 1 MPEG 2 [36] consists of five parts, and was issued as a Draft International Standard (DIS) in 1993. MPEG 2 [36] is capable of coding of standard television pictures at bitrates in the range from 3 Mbit/s to 15 Mbit/s. High definition TV pictures can be coded at bitrates between 15 Mbit/s and 30 Mbit/s.

The MPEG 4 [37] standard supports a wider range of applications than MPEG 1 [35] and MPEG 2 [36]. The video part of MPEG 4 includes both coding of live video and synthetic video. The live video part of the MPEG 4 coding algorithm is similar to ITU-T Recommendation H.263 [43]. MPEG 4 [37] consists for the time being of 11 parts. Part 2 of the standard defines the video coding. A text identical to ITU-T Recommendation H.264 [44] has recently been approved as part 10 of MPEG 4. A paper by Koenen [45] presents an overview of the basics of the initial MPEG 4 standard.

There are proprietary vide coding algorithms being frequently used for streaming applications. The most popular algorithms are:

- Real Video of Real Networks.
- Windows Media of Microsoft.

The codec used in Windows Media, VC-1, has recently been standardized by SMPTE, an organization developing industry standards for the imaging industry [54]. A paper by Bennet and Bock [55] considers video compression techniques developed since the ratifications of MPEG 2 [36] with a focus on MPEG 4/AVC (H.264/AVC [44] and VC-1 (Windows Media Video 9) [54]. The paper concludes that comparing these two codecs there is very little performance difference between them.

5.3.2 Video frame rate

The video frame rate for standard TV pictures is 25 frames per second (Europe) or 30 frames per second (North America).

Experiments carried out in the Vis-à-Vis project [46] has documented that there is a significant user perceived quality difference between 12,5 frames per second and 25 frames per second. Other experiments have indicated a threshold around 15-16 frames per second. It is worth noting that a number of products designed for communication on one or two B-channels on ISDN, support a frame rate of 10-15 frames per second, depending on the amount of changes in the picture, while equipment connected to a 384 kbit/s may support a frame rate of 25/30 frames per second.

A paper by Yadavalli et al. [47] reports the results of tests to determine the preferred frame rate at a fixed bit rate for low bit rate video. At a given bit rate, a lower frame rate the quality of individual frames are improved at the expenses of the smooth motion of the video. The tests reported were carried out at the different frame rates; 10, 15 and 30 frames per second. The test included low motion, medium motion and high motion sequences. Three video coding algorithms were used:

- Sorenson professional video coder version 2.1.
- H.263+ coder.
- A vawlet-based rate-distortion optimized coder developed at Cornell University, N.Y., USA.

In order to maintain similar quality across the three motion categories, the low motion sequences were encoded at 100 kbit/s, the medium motion sequences at 200 bit/s and the high motion sequences.

The evaluation was performed using the double stimulus, five-category scale of ITU-R Recommendation BT.500 [48]. For analysis purposes the ratings were converted into ordinal rankings.

The results show that there is a slight preference for 10 frames per second for the low motion sequence containing a news reader. Sequence of 15 frames per second was ranked highest for all other sequences. For low motion sequences 10 frames per second was preferred to 30 frames per second. For high motion there is a slight preference for 10 frames per second compared to 30 frames per second.

The 15 frames per second preference was valid for all three codecs used in the tests.

It is worth mentioning that the bit rate limitation may influence these results. Without such limitation or with higher bitrate there might be a stronger preference for higher frame rates. However, bit rate restrictions may be a realistic scenario for a large number of users.

Another experiment carried out by Apteker et al. [56] used the term *watchability* to describe the user acceptance of a video message. Watchability is a composite term that includes various aspects of user acceptance of a video message including the general relationship between visual and auditory components. A Video Classification Scheme (VCS) was developed. The scheme uses three dimensions:

- The temporal characteristics (T).
- The auditory characteristics (A).
- The visual characteristics (V).

Each characteristic was classified "high" or "low". These characteristics were used to classify the video clips used in the tests. The test persons were asked to rate each clip on a seven point Likert scale. The original video frame rate was 30 frames per second. The tests were carried out using 5, 10 and 15 frames per second.

The experiment was carried out in a multitasking environment where the primary task was to carry out spell checking on a Word document while monitoring the video as a secondary task. The third task was to rate the particular video clip using an on-line tool.

The mean results of the tests are presented in table 10. For 30 frames per second, the highest score (i.e. 7) is assumed.

One of the strong conclusions that can be made is the significant watchability difference between the three frame rate conditions tested. The authors characterize 15 frames per second as just acceptable, 10 frames per second is much less so, and 5 frames per second is very unacceptable. It is also concluded that the factors T, A and V influence perceptions and ratings of whether reduces frame rates are acceptable. An example is good audio and video characteristics in combination with low temporal video that is more affected by lower frame rate than the same characteristics in combination with high temporal video.

Table 10: Mean ratings for each combination of characteristics and frame rate

VCS classification	30 frames per second	15 frames per second	10 frames per second	5 frames per second
$T_{lo}A_{lo}V_{lo}$	7,0	6,4	6,2	6,3
$T_{lo}A_{lo}V_{hi}$	7,0	5,1	4,8	3,3
$T_{lo}A_{hi}V_{lo}$	7,0	5,9	5,2	3,5
$T_{lo}A_{hi}V_{hi}$	7,0	5,7	5,1	3,0
$T_{hi}A_{lo}V_{lo}$	7,0	6,5	6,3	5,9
$T_{hi}A_{lo}V_{hi}$	7,0	6,0	5,6	4,2
$T_{hi}A_{hi}V_{lo}$	7,0	6,4	5,9	4,9
$T_{hi}A_{hi}V_{hi}$	7,0	6,0	5,7	4,5

5.3.3 Video picture resolution

The ITU-T video coding standards (H.261 [42], H.263 [43], and H.264 [44]) use the term Common Intermediate Format (CIF) to define picture resolution. A CIF picture has 352 x 288 pixels for luminance.

ITU-T Recommendation H.261 [42] defines two formats; QCIF (176 x 114 pixels) and CIF. QCIF is mandatory in ITU-T Recommendations for multimedia systems. However, most implementations on the market today also support CIF.

ITU-T Recommendation H.263 [43] is a standard that may give a better picture quality than ITU-T Recommendation H.261 [42] at the same bitrate. As a rule of thumb the ITU-T Recommendation H.263 [43] quality at a bitrate is equal to the ITU-T Recommendation H.261 [42] quality at the double bitrate. Furthermore, ITU-T Recommendation H.263 [43] supports a wider range of bitrates than ITU-T Recommendation H.261 [42], from approx. 9,6 kbit/s to 2 Mbit/s.

ITU-T Recommendation H.263 [43] defines the following picture formats:

- Sub-QCIF (128 x 96 pixels).
- QCIF (176 x 114 pixels).
- CIF (352 x 288 pixels).
- 4CIF (704 x 576 pixels).
- 16CIF (1408 x 1 152 pixels).

The latest ITU-T Recommendation H.264 [44], supports most of the relevant picture formats including High Definition TV formats.

CIF is the picture format used for live video in most real-time applications for the time being. Systems using small screens may use QCIF or subQCIF. There are systems on the market that supports 4CIF for applications such as document camera. However, for the time being the processor capacity of the video processing DSPs are no high enough to support live video at 4CIF.

The MPEG 1 [35] and MPEG 2 [36] standards are defining a slightly different format than the ITU-T Recommendations, (SIF - 352 x 288/240 pixels). MPEG 2 [36] also supports the format defined in ITU-R Recommendation BT.601 [8] (720 x 576/480 pixels).

The live video coding part of MPEG 4 [37] is similar to H.263 [43], and annex 10 to MPEG 4 is identical to ITU-T Recommendation H.264 [44].

The perceived quality is related to the screen size; while CIF is required to achieve acceptable picture resolution for e.g. a 17' screen, subQCIF may be sufficient for a 4' screen.

5.3.4 Delay

The delay introduced by a video codec is larger than the delay of codecs designed for interactive speech communication. There is no paper presenting measurement results on the delay of a video codec.

An estimate made by Thoma [61] is 260 msec. However, the delay is depending on the video frame rate. A rule of thumb for state-of-the-art video codecs indicated by a video codec manufacturer is

$$Video\,coding\,delay\,(ms) = 3\frac{1\,000}{Video\ frame\,rate} \tag{5}$$

5.3.5 Overall rating

PSNR (Peak Signal-to-Noise Ratio) is a frequently used video quality indicator. However, there is no reliable relation between PSNR and user perceived video quality. The ITU-R/ITU-T Video Quality Experts Group (VQEG) has evaluated objective models for assessing user perceived video quality. The work has been organized in two phases. There are reports available for each of the phases [57] and [58].

Phase II contains two groups of subjective tests, one for 525-line video and one for 625-line video. Each group spans a wide range of quality, so that the evaluation criteria are able to determine statistical differences in model performance. The results of the tests are given in terms of Differential Mean Opinion Score (DMOS). These results were compared with the results of calculation algorithms proposed by different organizations. For 525 line video two algorithms performed better than the other. For 525 line video 4 algorithms performed better than the other.
Based on this work both ITU-R and ITU-T have adopted four models for standardization [59] and [60].

There are three basic methods to perform objective measurements:

- FR: A method applicable when the full reference video signal is available. This is a double ended method and is the subject of ITU-T Recommendation BS.1387 [41].
- RR: A method applicable when only reduced video reference information is available. This is also a double-ended method and is the subject of a separate Recommendation.
- NR: A method applicable when no reference video signal or information is available. This is a single-ended method and the subject of a separate Recommendation.

The present Recommendations describe the full reference method. Models for the other methods are for further study.

Figure 9 illustrates a Full Reference (FR) perceptual quality measurement set-up.

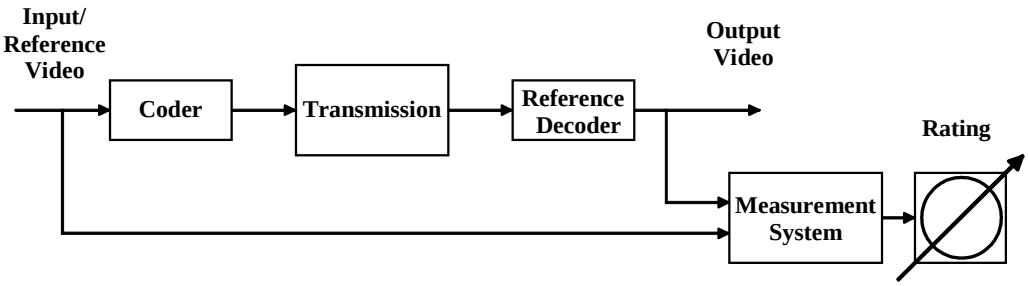


Figure 9: Full reference (FR) perceptual quality measurement set-up

Further details about the methods are given in annexes to the Recommendations [59] and [60].

5.3.6 Packet loss

Hayashi et al. [62] have evaluated the effects of packet loss on the subjective quality of MPEG1 video. Two signal sequences were used when testing:

- A news reader (SS-1).
- People dancing to rock music (SS-2).

The tests were carried out with video packet loss ratios from 0 % to 5 %. The accompanying audio was not degraded.

The picture frame rate was changed from 2 frames per second to 30 frames per second. However, packet loss was only added at 10 frames per second and 30 frames per second video. The video transmission bit rate was 800 kbit/s (30 frames per second) or approximately 550 kbit/s (10 frames per second). The test method and scale used conformed to ITU-R Recommendation BT.500 [48].

The results of these tests are illustrated in figure 10. The high motion sequence (SS-2) is more sensitive to packet loss than the low motion sequence (SS-1). For 2 % and 5 % packet loss 10 frames per second was rated higher than 30 frames per second irrespective of the sequence content. It can also be concluded that to obtain acceptable quality (i.e. MOS equal to 3 or higher) of the video content, the packet loss should not exceed 1 %.

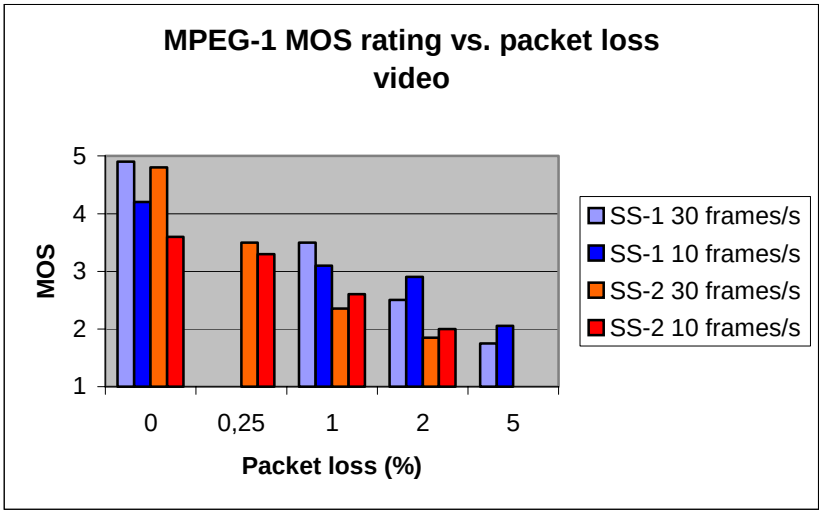


Figure 10: MPEG1 video. Effects of packet loss on subjective quality

5.3.7 Bit Rate

As a rule of thumb the ITU-T Recommendation H.263 [43] quality at a bitrate is equal to the ITU-T Recommendation H.261 [42] quality at the double bitrate, in other words, the bitrate may be reduced by 50 %.

The overall performance of the ITU-T Recommendation H.264 [44] algorithm is as such that bit rate savings of 50 % or more, compared to the current state of technology (ITU-T Recommendation H.263 [43]), are reported.

5.4 End-to-end performance characteristics for text, data and image service components

5.4.1 Reliable transport protocols

5.4.1.1 Transport Control Protocol (TCP)

Most non-real time IP communication applications use the TCP/IP Protocol suite for data transport. TCP [63] is a protocol that allows devices to establish and manage connections and send data reliably, and takes care of handling potential problems that can occur during transmission. TCP is stream-oriented; it is designed to have applications send data to it as a stream of bytes, rather than requiring fixed-size messages to be used. This provide maximum flexibility for a wide variety of uses, because applications do not need to worry about data packaging, and can send files or messages of any size. TCP takes care of packaging these bytes into messages called segments. Furthermore, TCP is a reliable protocol, when a packet is lost, it is retransmitted. This mechanism is based on allocating a sequence number to each packet sent. The receiver acknowledges the receipt of packets. If a packet is not acknowledged after a period of time, it is retransmitted by its sender. The period of time is usually determined by the Round Trip Time (RTT) of the connection. Retransmissions or other events may also influence this period.

One problem with this basic scheme is that the transmitter cannot send a second message until the first has been acknowledged. To cope with this problem the concept of "Sliding Window" has been introduced. The "Sliding Window" is the maximum number of bytes the receiver can accept. This Window limits the number of unacknowledged data transmitted. It is defined in the TCP header. The transmitter may transmit all packets within the window without receiving an acknowledgement.

5.4.1.2 Stream Control Transmission Protocol (SCTP)

Like TCP, SCTP is a reliable transport protocol used in an IP environment. SCTP is defined in RFC 2960 [64]. The initial application of SCTP was transport of signalling data, and it is used by SIGTRAN protocols. The most significant difference between SCTP and TCP is that a SCPT stream is a sequence of messages, while a TCP stream is a sequence of bytes. Furthermore, SCTP supports multihoming and several streams within a connection.

RFC 3286 [65] presents a high level introduction to the capabilities supported by the Stream Control Transmission Protocol (SCTP).

5.4.2 Reliable transport protocol performance

The performance of TCP is influenced by the Round Trip Time (RTT) of the connection as well as packet loss probability and any bottlenecks.

There are a number of papers discussing the performance of TCP, e.g. T.V. Lakshman and Madhow [66]. It is concluded in the paper that random packet loss leads to significant throughput reduction when the product of the loss probability and the square of the bandwidth-delay product is larger than one. It is also observed that for high bandwidth-delay products TCP is unfair towards connections with high propagation delay. Other aspect addressed in the paper is the TCP window adjustment mechanism.

The paper also concludes that TCP's vulnerability to random loss makes it difficult to multiplex data traffic with real-time traffic with a rapidly time-varying rate (e.g. video traffic), especially if both kinds of traffic share the same buffer. This problem may of course be reduced when network QoS mechanisms are implemented.

It is also indicated that for guaranteed performance in highly utilized network, each TCP connection should be given reserved buffer and bandwidth resources throughout the network.

A paper by Mathis et al. [68] presents a simple formula describing the throughput of the TCP protocol in the event of random packet loss:

$$Throughput = C\frac{MSS}{RTT}\frac{1}{\sqrt{p}} \quad (6)$$

Where:

C is a constant.

MSS is the Maximum Segment Size (typically 1 460 Bytes).

RTT is the Round Trip Time.

P is the packet loss probability.

This simple formula has several limitations. The paper highlights the following:

1) The data receiver is announcing too small TCP window. In this case the TCP performance is likely to be controlled by the receiver's window.

2) If the sender does not always have data to send, the model is not likely to apply.

3) The elapsed time consumed by TCP timeouts is not modelled.

4) Not all TCP implementations are covered.

5) Short connections may not fit the model.

Padhye et al. [69] presents a more generic formula overcoming some of the limitations of the formula above. A simplified approximation of this formula is:

$$Throughput \approx \min\left\{\frac{W_{\max}}{RTT}, \frac{1}{RTT\sqrt{\frac{2bp}{3}} + T_0 \min\left(1, 3\sqrt{\frac{3bp}{8}}\right) p\left(1+32p^2\right)}\right\} \quad (7)$$

Where:

W_{max} is the maximum TCP Window of the application,

b is the number of packets that are acknowledged by an ACK message,

T_0 is the period of time the sender waits before a packet is retransmitted.

Another issue is the effects of asymmetrical access networks on the performance of TCP. This is discussed in a paper by Lakshman et al. [67]. The paper introduces a term, *normalized asymmetry*, which is defined as the ratio of the transmission time of ACKs on the bottleneck link to the transmission time of the packets on the other link.

To obtain good utilization of the forward link, the paper concludes that the forward link buffer size is at least equal to the normalized asymmetry. It is also concluded that asymmetry increases TCP's already high sensitivity to random loss anywhere in the forward path of the TCP connection.

5.4.3 User experience rating

J. Nielsen [70] indicate that three different perceptual response time regions are identified:

- **Instantaneous experience: 0,1 second** is about the limit for having the feel that the system is reacting instantaneously, an important limit for conversational services (e.g. chatting).

- **Uninterrupted experience: 1,0 second** is about the limit for the user's flow of thought to stay uninterrupted, even though the user does lose the feeling that the service is operating directly, an important limit for interactive services (e.g. gaming).

- **Focused experience: 10 seconds** is about the limit for keeping the user's attention focused on the dialogue. For longer delays, users will want to perform other tasks while waiting for the computer to finish, so they should be given feedback indicating when the computer expects to be done. Feedback during the delay is especially important if the response time is likely to be highly variable, since users will then not know what to expect.

For download times users tend to adapt their quality judgement towards the expected download time [71].

A network operator can be able to monitor and control network QoS parameters, while the user experience can be expressed as a QoE rating. An important issue is therefore the relationship between QoS parameters and QoE. The relationship between QoS and QoE for traditional Internet Services such as download of information (Web browsing) has been studied by Khirman and Henriksen [72]. They measured three characteristics,

a) The relationship between QoE and the time to receive the first byte of a response.

b) The relationship between QoE and the effective bandwidth.

c) The relationship between QoE and delivery time of an object.

The tests were carried out on real-life connections. About 80 % of the connections were set up via modems (up to 56 kbit/s); the remaining 20 % of the connection were from corporate LANs with high-speed access. The QoE metric was cancellation rate. The cancellation rate is defined as the number of http requests that were stopped divided by the total number of http requests.

The tests showed that for response times below 50 msec the cancellation rate increased from 2,3 % at 10 msec to 3,8 % at 50 msec. For response times above 50 msec there was almost no cancellation rate increase.

A strong cancellation rate dependency on the delivery bandwidth is shown. In the 15 kbit/s to 100 kbit/s delivery bandwidth range an estimate for the cancellation rate is given by the formula:

$$\text{Cancellation\_Rate} = -0{,}017 \times \ln(\text{Delivery\_bandwidth}) + 0{,}13 \quad (8)$$

The delivery bandwidth is given in kbit/s.

The cancellation rate as a function of delivery bandwidth is illustrated in figure 11.

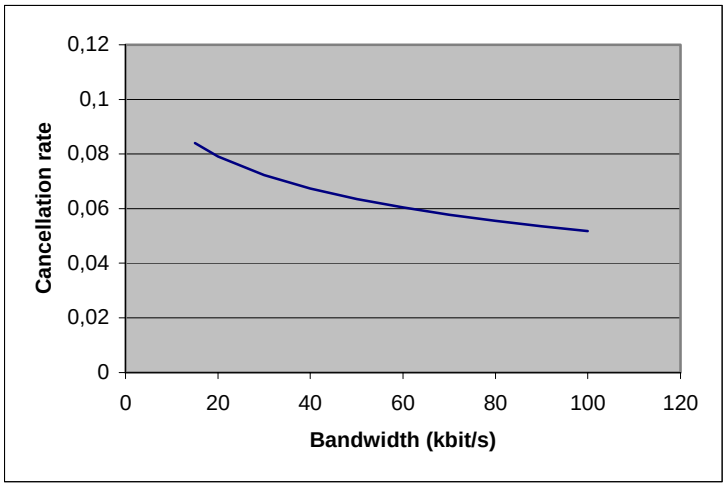


Figure 11: Cancellation rate as a function of bandwidth

When the delivery bandwidth increases from 100 kbit/s to 200 kbit/s there is a significant cancellation rate reduction from 5 % to 1 %. No significant customer satisfaction improvement was found when the delivery bandwidth was above 200 kbit/s.

Finally the relation between the cancellation rate and delivery time was analysed. For objects that were only partially delivered, an estimate of the delivery time of the whole object was used. For delivery time above 1 second there is a nearly linear relationship described by the formula:

$$\text{Cancellation\_rate} = 0{,}0006 \times \text{Delivery\_Time (seconds)} + 0{,}014 \quad (9)$$

The formula is illustrated in figure 12.

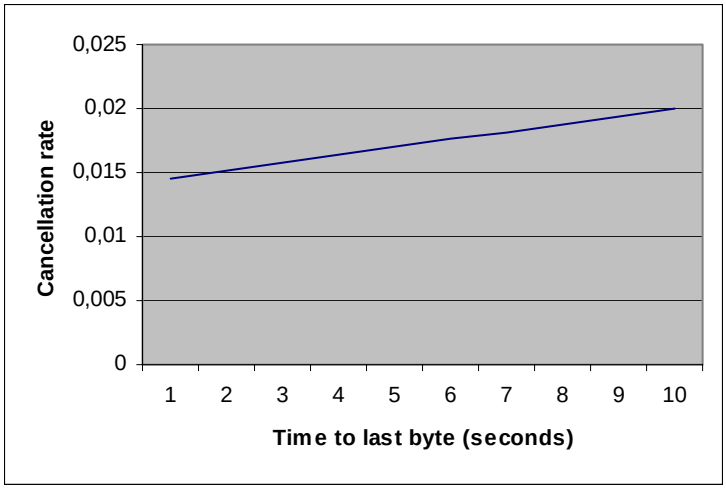


Figure 12: Cancellation rate as a function of delivery time

A new ITU-T Recommendation [73] describes a methodology for estimating end-to-end performance in IP networks for data applications. Annex B of the Recommendation present results from tests on web browsing user perceived quality.

Subjective experiments with maximum session times of around 6 seconds, 15 seconds and 60 seconds were set up A single experiment consists of 49 sessions, and both experts and naïve (untrained) users participated. The users rated the perceived quality as Mean Opinion Score (MOS) using the Absolute Category Rating Scale.

The experiments showed that the perceived quality goes down linearly with the logarithm of the Session Time for both expert and naïve users.

For browsing sessions with long session times (about 60 sec) the MOS can be predicted using the formula:

$$MOS = 5{,}72 - 0{,}936 \ln(SessionTime), \text{ clipped between MOS 1,0 and 5,0} \quad (10)$$

Browsing sessions with shorter session times showed lower correlations between session time and subjective quality. A model where it is assumed that the last download time has a more severe impact than the preceding sessions of the experiment is therefore introduced. The term *Weighted Session Time* where each of four intervals is given different weights is calculated. The last session is given more weight than the other sessions.

The weights are given in table B.1 of the draft Recommendation.

A formula has been developed for each of the session duration times:

$$MOS = 5{,}76 - 0{,}948 \ln(Weighted\ Session\ Time), \text{ clipped between MOS 1.0 and 5.0 for long duration sessions} \quad (11)$$

$$MOS = 4{,}79 - 1{,}03 \ln(Weighted\ Session\ Time), \text{ clipped between MOS 1.0 and 5.0 for medium duration sessions} \quad (12)$$

$$MOS = 4{,}38 - 1{,}30 \ln(Weighted\ Session\ Time), \text{ clipped between MOS 1.0 and 5.0 for short duration sessions} \quad (13)$$

5.5 Media quality interaction

5.5.1 Lip synchronization

The video coding introduces larger delay than the speech coding. Most of the video packets are larger than the speech packets, which may results in a slightly larger network delay, particularly in the access network. The result is than the speaking motion of a person is not synchronized with the speech.

ETR 297 [75] specifies that to preserve lip synchronization the delay difference between speech and moving image should not exceed:

- 40 msec when speech arrives after moving image.
- 20 msec when speech arrives before moving image.

ANSI T1.552 [14] specifies that the perception differences are the values specified in the ETSI document, while the acceptable differences are 80 msec before the moving image to 80 msec after the moving image. These values may be based on experiments reported by Steinmetz [74].

There are studies indicating a larger range. Summerfield [76] suggests that integration of audio and visual channels can occur up to around a lag of 80 msec. Jardetzky et al [77] have shown that mismatch between audio and video streams can go up to 80 msec to 100 msec before lack of synchronization is detected. Longer delays can be tolerated if the audio channel lags behind the visual channel. Studies by Lewkowicz [78] suggest that when audio precedes video, asynchronies can be detected with delays in the region of 65 msec to 80 msec. When video precedes audio, this threshold rises to 112 msec to 140 msec.

For broadcasting purposes ITU-R Recommendation BT.1359-1 [79] defines detectability and acceptability threshold for lip synchronization. Figure 13 describes these thresholds. The detectability thresholds are 125 msec when sound is delayed with respect to the video, and 45 msec when sound is advanced with respect to the video. The acceptability thresholds are 185 msec when sound is delayed with respect to the video, and 90 msec when sound is advanced with respect to the video.

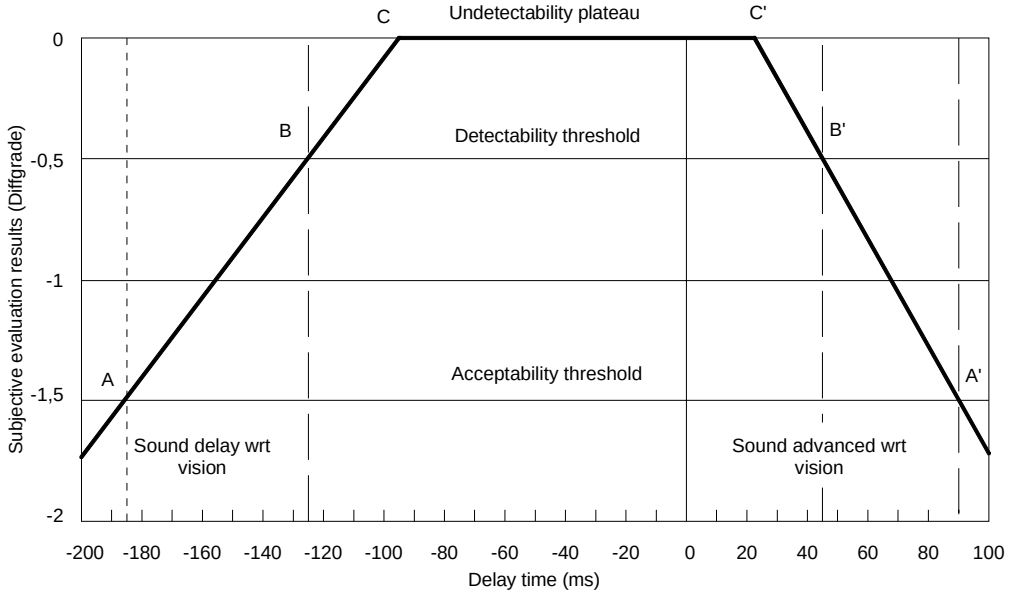


Figure 13: Detectability and acceptability thresholds for lip synchronization defined in ITU-R Recommendation BT.1359-1 [79]

These publications indicate that the threshold is not symmetric; an audio channel that lags behind the visual channel is less annoying than the opposite.

Although the information above indicates that some asynchrony is not perceived or might be accepted, it is likely that the speech has to be delayed to meet the objectives indicated. However, the delayed speech may result in a lower perceived MM QoS than unsynchronized media where no extra delay is added to the speech.

It is not clear how this extra delay affects the perceived voice quality; lip synchronization has a positive effect while extra delay has a negative effect. Further considerations and tests are required.

5.5.2 Multimedia class of service

The demanding part of defining the MM CoS is to assess the interaction effects between audio and video. Tests reported by Hollier and Voelcker [80] and Hollier et al. [81] have shown strong intermodal dependency between audio and video. Although the interpretation is uncertain, it has been observed that when no video is present, the perceived audio quality is rated lower than if video is present. It has also been observed that unchanged audio is rated lower when the video quality is low compared to the rating when the video quality is higher. Another observation made is that the cross-modal interaction is greater where the test material comprises talking head content.

The influence of video quality on perceived audio quality and vice versa is addressed in a paper by Beerends and de Caluwe [82]. In the experiments subjects were asked to judge the quality of two television commercials. Five different modes were used in the experiments,

1) An audio only mode where the subjects only heard the audio tracks and rated the audio quality.

2) An audiovisual mode where the subjects were given the combined (audio and video) stimulus, but were asked to rate the audio quality only.

3) A video only mode where the subjects only saw the video and were asked to rate the video quality.

4) An audiovisual mode where the subjects were given the combined (audio and video) stimulus, but were asked to rate the video quality only.

5) An audiovisual mode where the subjects were given the combined (audio and video) stimulus, and were asked to rate the overall audiovisual quality.

Commercial loudspeakers were used for audio presentation, and the viewing conditions were based on ITU-R Recommendation BT.500 [48].

Four audio quality levels and four video quality levels were included in the tests. The highest quality is CD quality (audio) and full resolution TV (ITU-R Recommendation BT.601 [8]). The audio signal was bandwidth limited where the lowest quality is the telephone bandwidth (300 Hz to 3 400 Hz). The video signal was degraded using an impairment system that is an extension of a system described in Appendix I of ITU-T Recommendation P.930 [83].

The user perceived quality was assessed using a nine-point Absolute Category Rating (ACR) scale.

The results analysis shows that there is a significant influence of audio quality on video quality and vice versa. The analysis also shows that the influence of video on audio is stronger than the reverse. The results also show that the video quality dominate the overall audiovisual quality. The paper presents the following data model mapping of audio and video quality to audiovisual quality,

$$MOS_{AV} = 1{,}12 + 0{,}007 MOS_A + 0{,}24 MOS_V + 0{,}088 (MOS_A \times MOS_V) \quad (14)$$

Where:

MOS_{AV} is the audiovisual rating,

MOS_A is the audio only rating,

MOS_V is the video only rating.

Another set of experiments is described in a paper by Hands [84]. Two experiments were carried out.

The first experiment compared the test sequence with a reference (un-degraded) sequence. This method is called double stimulus continuous quality scale (DSCQS). A five point rating scale was used. The subject moved a pointer along the scale to indicate their rating, and rated both the test sequence and the reference sequence. The source material for both presentations is identical in content. Subjects were not informed which of the two sequences is the test and which is the reference, and the sequence is randomized. The difference score was computed.

The test material consists of short (5 sec) head and shoulder sequences. The subjects rated:

- Audio quality.

- Video quality.
- Overall quality.

When the audio was rated, there was no video signal, and when the video was rated there was no audio signal.

The second experiment consists of two sets of test material:

- A head and shoulder sequence.
- A high motion sequence.

Throughout this experiment the single stimulus quality scale (SSQS) methodology was used. The subjects reported the rating verbally using a five-point scale.

For both experiments the audio was degraded by using a Modulated Noise Reference Unit. The video was degraded by manipulating the degree of blockiness present in the video (i.e. changing the quantization of the coded picture).

The analysis of the results of the first experiment gave the following formula for audiovisual quality estimation:

$$Multimedia\,Quality = 0{,}85\,Audio\,Quality + 0{,}76 Video\,Quality - 0{,}01(Audio\,Quality \times Video\,Quality) - 3{,}34 \quad (15)$$

The analysis of the second experiment gave different formulas for the head and shoulder sequence and the high motion sequence. The regression analysis of the head and shoulder data gave best predictive power (R^2) for the following single multiplicative term formula:

$$Multimedia\,Quality = 0{,}17(Audio\,Quality * Video\ Quality) + 1{,}15 \quad (16)$$

The analysis of the high motion sequences gave the following formula:

$$Multimedia\,Quality = 0{,}25\,Video\,Quality + 0{,}15(Audio\,Quality \times Video\,Quality) + 0{,}95 \quad (17)$$

The paper [84] concludes that:

1) The form of the predictive model is content dependent.
2) Subjective quality ratings are formulated using a multiplicative rule.

Two papers by Winkler and Faller [85], [86] present the results of subjective audio, video and audiovisual quality assessment in scenarios that are typical for mobile applications. The test material that was originally in TV format, was downsampled to QCIF. The video coding algorithms used were:

- MPEG 4 (Part 2) [37].
- ITU-T Recommendation H.263 [43].
- ITU-T Recommendation H.264 [44].

The video frame rate was reduced to 8 frames per second or 15 frames per second.

The audio codec used was conforming to the MPEG-4 AAC-LC [37] coding standard.

Like the results presented by Hands [84], formulas for estimation the audiovisual quality from the audio and video quality ratings may be based a multiplicative term. The single multiplicative term formula is:

$$MOS_{AV} = 1{,}98 + 0{,}103(MOS_A \times MOS_V) \quad (18)$$

However, the additive linear model:

$$MOS_{AV} = -1{,}51 + 0{,}456 MOS_A + 0{,}770 MOS_V \quad (19)$$

gave a somewhat better fit.

A paper by Kitawaki et al. [87] describes an objective perceptual multimedia quality model for audiovisual communications, taking account of the mutual interaction of audio and video information. The primary use of the model is to measure the quality of limited bandwidth services.

Instead of using the audio and video quality being independent of each other, the model takes into account the influence of audio when accessing video an vice versa. The paper identifies these characteristics as *Mutual interaction quality.*

The multimedia quality is estimated as follows:

1) The audio quality and video quality in multimedia is estimated from objective measurements of audio and video.

2) The multimedia quality is calculated.

The paper presents the results from subjective tests carried out to verify the model. In the subjective tests the ACR (Absolute Category Rating) method according to ITU-T Recommendation P.911 [6] was used.

Media were encoded using the MPEG-4 video coder. The video frame rate was 15 frames/s and the picture resolution was 320x240 pixels. The video bitrate was 120 kbit/s, 320 kbit/s, 512 kbit/s and reference.

The audio was encoded using the MPEG-4 AAC audio coder using 48 kHz sampling. The audio bitrate 32 kbit/s, 48 kbit/, 64 kbit/s and reference. The audio signal was stereo.

Three source signals used were reports, news, scene + BGM.

The characteristics assessed are described in table 11.

Table 11: Content characteristics used in subjective assessment

	Source signal	Quality factor	Score
1	Audio and Video	Multimedia quality	MOS_{AV}
2	Audio and Video	Mutual interaction quality: audio quality and video quality in multimedia	$A_q(V_q)$ $V_q(A_q)$
3	Video	Media independent quality: video quality	A_q
4	Audio	Media independent quality: audio quality	V_q

The multimedia quality could be estimated using the formula:

$$MOS_{AV} = 0{,}188 V_q(A_q) + 0{,}211 A_q(V_q) + 0{,}112 V_q(A_q) \times A_q(V_q) + 0{,}618 \quad (20)$$

Where:

$A_q(V_q)$ is the audio mutual interaction quality,

$V_q(A_q)$ is the video mutual interaction quality.

Formulas for estimating $A_q(V_q)$ and $V_q(A_q)$ are:

$$A_q(V_q) = 0{,}890 A_q + 0{,}043 V_q + 0{,}387 \quad (21)$$

$$V_q(A_q) = 0{,}039 A_q + 0{,}931 V_q + 0{,}099 \quad (22)$$

By combining the formulas above it is possible to develop a formula that can be compared with the formulas presented in [82], [84], [85] and [86].

$$MOS_{AV} = 0{,}000429 A_q^2 + 0{,}004484 V_q^2 + 0{,}206681 A_q + 0{,}224931 V_q + 0{,}092990 A_q V_q + 0{,}727560 \quad (23)$$

6 Conclusions

The information presented shows that there are several factors that influence the quality of each multimedia service component and of a multimedia service as a whole.

It is also shown that techniques exist that may reduce or remove the effects of some network related degradations. The best example is packet loss, which can be repaired by retransmission or by packet loss concealment. However, there might be negative side effects, e.g. increased delay when retransmission is used.

To assess the performance of each service component as well as the performance of a multimedia application, subjective tests may be carried out.

There are also available tools for estimating the performance of service components. However, the grade of accuracy varies, narrowband speech is the service component where the most accurate estimate can be made.

To estimate the effects of interactions between service components is a challenge. There are differences between the results presented addressing the interaction between audio and video. There are also indications that the interactions are content dependent.

There is no information on the effect of the quality of a data service component (e.g. web browsing) of a multimedia service.

History

Document history		
V1.1.1	February 2006	Publication