# PACKET-LAYER MODEL FOR QUALITY ASSESSMENT OF ENCRYPTED VIDEO IN IPTV SERVICES

Qian Zhang*, Ning Liao†, Fan Zhang‡, Zhibo Chen§ and Lin Ma¶
*School of Information and Control Engineering , Xi'an University of Architecture and Technology, Xi'an, China
Media Processing Laboratory, Technicolor Research and Innovation, Beijing. E-mail:qzhangmis@gmail.com
†Media Processing Laboratory, Technicolor Research and Innovation, Beijing, China. E-mail:ningliao@technicolor.com
‡Lenovo Cooperation Research, Hong Kong
Media Processing Laboratory, Technicolor Research and Innovation. E-mail:zhangfanhk@lenovo.com
§Media Processing Laboratory, Technicolor Research and Innovation, Beijing, China. E-mail:zhibochen@technicolor.com
¶Huawei Noahs Ark Lab, Hong Kong. E-mail:lma@ee.cuhk.edu.hk

***Abstract*—In this paper, a packet-layer quality assessment model is proposed for predicting the subjective quality of encrypted video for IPTV services. The detected information from encrypted video-stream is parsed into frame layer, and a novel estimation of frame type and Group-Of-Picture (GOP) structures is proposed to assist the parameter extraction utilized in the model. An efficient loss-related parameter is developed to reveal the visible degradation by loss. The quality assessment model focuses on predicting the quality measurement caused by both coding and channel artifacts. The cross-validation results on numerous databases show that the proposed model is not only better than other compared ones in performance, but also more generalized and robust to various testing conditions.**

## I. INTRODUCTION

With the development of IP networks, video communication over wired and wireless IP networks (for example, IPTV service) has become popular. Unlike traditional video transmission over cable networks, video delivery over IP networks is less reliable. Consequently, in addition to the quality loss from video compression, the video quality is further degraded when a video is transmitted through IP networks. To make the IPTV services meet the high expectation of the end-users, predicting and monitoring the quality of services (QoS) and the users' quality of experience (QoE) are indispensible for quality design and management. A successful video quality modeling tool needs to rate the quality degradation caused by network transmission impairment (for example, packet losses, transmission delays, and transmission jitters), in addition to quality degradation caused by video compression.

The standard group ITU-T SG 12 has been devoted for standardized recommendations (G. 107 [1], G. 1070 [2]) for network quality-planning, and models (P.NAMS [5], P.NBAMS [6]) for in-services quality monitoring. Based on the input information and primary applications, the objective quality assessment method can be categorized into five types, which are the media-layer model, parameter packet-layer model, parametric planning model, bit-stream layer model and hybrid model [7]. As the payload information is usually encrypted for example, in IPTV, the bit-stream model (like P.NBAMS) cannot be applied at the device where the encrypted bit-stream cannot be de-encrypted. The packet-layer model (like P.NAMS) can be applied to estimate the perceived video quality by solely using the packet-header information. As a consequence, the measurement of packet-layer model is lightweight without accessing to the media signal itself. Additionally, it is applicable when the processing load is encrypted or very limited, e.g. monitoring the QoE inside set-top box (STB).

In the related literatures considering the packet-loss degradation, packet-loss rate (Ppl) [2], packet-loss frequency (PLF) [3], eXtended Weight per SEQuence (xwpSEQ) [4], MLoVA [8] are extracted as representative metrics and used for modeling the packet loss extent and its impact on quality. Ppl and PLF are non-frame-layer metrics calculated with respective to a sequence sample, which do not distinguish the loss impact in detail. xwpSEQ and MLoVA considered the loss with frame-layer information. However xwpSEQ's performance is still limited due to poor modeling of the visibility of artifacts, and MLoVA only handle the packet loss degradation.

In our model, an efficient loss-related parameter is proposed based on frame-layer information by a novel estimation, which can better reveal the visibility of artifacts and predicted quality. The technical challenges we solved and the differentiators of our model with others' lie in the following four aspects: (1) Different number of slice per frame is taken into account and scaled in a more efficient way to predict the quality. (2) Frames are classified into four types by a novel estimation, and different weights are assigned accordingly. (3) Coding and channel artifacts are modeled simultaneously using bitrate and proposed loss-related parameters. (4) The model is capable of handling either fixed or adaptive GOP length from different encoder configurations.

The remainder of the paper is presented as follow. The proposed packet-layer quality assessment model is described in Section II. The experimental setting and results is shown in

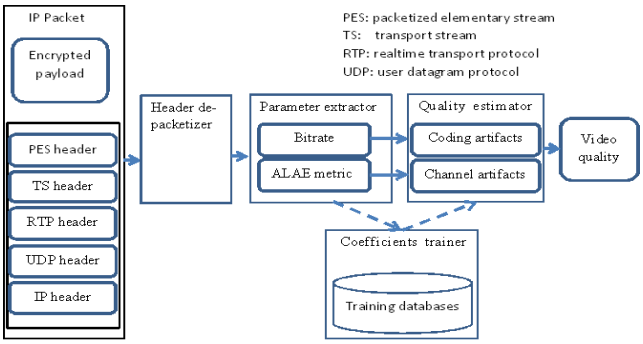



Fig. 1. Framework of the quality assessment model.

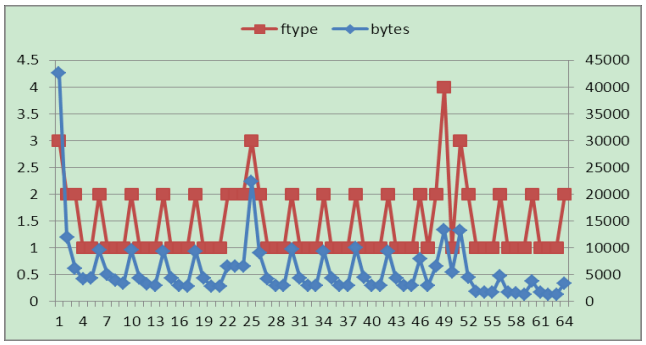



Fig. 2. Frame type estimation: x-axis is frame number. Left y-axis is frame type and right y-axis is bytes.

Section III. In Section IV, we summarize our model with its advantages.

## II. QUALITY ASSESSMENT MODEL

### A. Framework

This paper proposes a packet-layer model on the purpose of predicting the quality of encrypted video with higher accuracy, better robustness and lighter computational load. As depicted in the model framework in Fig. 1, the input is the encrypted payload with packet headers, and the output is the video quality. The model designed between input and output has four units: header de-packetizer is to de-packetize the stream and parse the header information; parameter extractor is to extract the parameters using frame-layer information; coefficients trainer is to train coefficients in the quality prediction model; quality estimator is to predict the video quality with trained coefficients.

### B. Frame type estimation

The bitstream is first de-packetized and header information is parsed into frame-layer parameters, e.g. bytes, Ppl. Losses happening in different types of frames with different levels of spatial complexity may result in different levels of visible artifacts, which lead to different quality measurement from subjects. For example, the effect of loss occurring in a reference I or P frame is more severe than that in a non-reference B frame. Frame type is estimated based on an estimated GOP structure and the number of bytes in a frame.

We define four frame types $ftype$ = {4 (scene-cut frame), 3 (non scene-cut I frame), 2 (P frame), 1 (B frame)}. An Intra frame can be determined from a syntax element, for example, "`random_access_indicator`" in the adaptation field of transport stream (TS) packet. A scene-cut frame is estimated as a frame that scene cut may happen and thus has highest spatial complexity. Considering different implementation of encoder with different types of GOP structure, a scene-cut frame may occur at an Intra frame or a non-Intra frame. For a bitstream with an adaptive GOP structure, scene-cut frames mainly correspond to intra frames with quite short GOP length in (1b). For a bitstream with a fixed GOP length, scene-cut frames may be non-Intra frames with quite large numbers of bytes in (1a). We define the frame $i$ as the scene-cut frame using following equation:

$$ftype_i = 4\ if \begin{cases} bytes_i > PRE_{IBytes} & \&i \in nonIntraframe \quad (1a) \\ glen_j < 0.5 * AVE_{GOPL} & \&i \in Intraframe \quad (1b) \end{cases}$$

where $bytes_i$ is the bytes of frame $i$, $PRE_{IBytes}$ is the bytes of previous I frame. $glen_j$ is the GOP length of GOP $j$ containing frame $i$, and $AVE_{GOPL}$ is the average GOP length. A GOP starts from scene-cut or non scene-cut I frame till the next scene-cut or non scene-cut I frame.

$$ftype_i = \begin{cases} 2, if\ bytes_i > AVE_bytes_j & (2a) \\ 1, if\ bytes_i \leq AVE_bytes_j & (2b) \end{cases}$$

To decide whether a non-Intra frame $i (i \in GOP_j)$ is P or B frame in (2), $AVE_bytes_j$ is calculated as the average bytes of GOP $j$ by excluding the scene-cut or non scene-cut I frame. If $bytes_i$ is larger than $AVE_bytes_j$ in (2a), frame $i$ is P frame, and is B frame otherwise in (2b). An example of frame type estimation is shown in Fig. 2.

### C. Artifact level estimation

Inspired by the work in [8], we propose a loss-related metric named **Averaged Loss Artifact Extension** (ALAE) to measure the visible degradation caused by video transmission loss. For each frame $i$, the Loss Artifact Extension (LAE) is calculated as the sum of Initial Artifact (IA) caused by the loss in the current frame and Propagated Artifact (PA) caused by the loss in reference frames:

$$LAE_i = IA_i + PA_i \quad (3)$$

$IA_i$ is calculated by:

$$IA_i = w_i^{IA} \times \frac{lp_i}{tp_i} \quad (4)$$

where $lp_i$ is the number of lost packets (including packets lost due to unreliable transmission and the packets ensuing the lost packets in current frame) and $tp_i$ is the number of total packets (including the estimate number of lost packets). $w_i^{IA}$ is the weight of initial artifact indicating the spatial complexity of loss and being dependent on the frame type, which is set as Table. I. Because loss occurred in a scene-cut frame often causes most serious visible artifacts for observers, its weight is set to be the largest. Non scene-cut I frame and P frame

TABLE I
IAE WEIGHT BASED ON FRAME TYPE.

| Frame type | Scene-cut frame | I frame | P frame | B frame |
|---|---|---|---|---|
| $w_i^{IA}$ | 1.0 | 0.3 | 0.3 | 0.01 |

usually cause similar visible artifacts since they are both used as reference frames, so the weight are set to be the same. Non-reference B frame is with the smallest weight.

$PA_i$ equals:

$$PA_i = w_i^{PA} \times ((1–\alpha) \times LAE_{pre1} + \alpha \times LAE_{pre2}) \quad (5)$$

where $(1–\alpha) \times LAE_{pre1} + \alpha \times LAE_{pre2}$ is the propagated error from its two previous reference frames. $\alpha$ is set to 0.25 for P frame and 0.5 for B frame. $w_i^{PA}$ is the weight of propagation artifact. It is 1 for P and B frame which means no artifacts attenuation, and 0.5 for loss-occurred Intra frame (regardless whether it is a scene-cut frame or not) which means the artifacts is attenuated by half. If Intra frame is successfully received without loss, $w_i^{PA}$ is set to 0 which means no error propagation.

Finally, the ALAE is calculated by:

$$ALAE = (\frac{1}{N} \sum_{i=1}^{N} LAE_i)/(f \times g(s)) \quad (6)$$

where $N$ is the number of frame and $f$ is the framerate. $g(.)$ is the function of $s$ and $s$ is the number of slice per frame. It should be noted that one frame may be encoded into several slices and each slice is an independent decoding unit. The number of slices in a frame impacts the video quality, and it is considered in quality modeling. However, in the encrypted bitstream, slices within a frame cannot be partitioned only using header information, and the location of a lost packet in the slice is unknown. The impact of $s$ on estimating the video quality is determined by the function $g(s) = \sqrt{s}$ from training databases.

### D. Overall quality prediction model

A video program may be compressed into various coding bit-rates, thus with different quality degradation due to video compression. The quality prediction model is capable of predicting the video quality combining the coding artifacts with channel artifacts, which can be obtained by a logistic function:

$$V_q^N = (\frac{1}{1 + a \times Br^b \times ALAE^c}) \quad (7)$$

In 7, one parameter $Br$ (bitrate) is used to model coding artifacts and $ALAE$ is used to model slicing channel artifacts. $a$, $b$, $c$ are constants obtained from curve-fitting using a least-square fitting method through 5 training databases. $V_q^N$ is the Normalized Mean Opinion Score (NMOS) within [0,1]. It should be noticed that the overall model can be reduced to predict the coding degradation only when $c = 0$.

TABLE II
DATABASE CONFIGURATION: DF-DISPLAY FORMAT(P-PROGRESSIVE; I-INTERLACE); BR-BITRATE(MBPS); FR-FRAMERATE(FPS); NS-NO. OF SLICE PER FRAME.

| | Training | Validation |
|---|---|---|
| Df | 1080p/i, 720p, 576i | 1080p/i, 720p, 576i, 480i |
| Br | 15, 9, 7, 6, 2.5, 2, 1, 0.5 | 15, 7, 6, 5, 4, 3.5, 3, 2.5, 2, 1.5, 0.5 |
| Fr | 50,30,25 | 60,50,30,25 |
| Ns | 1, 18, 68 | 1, 15, 18, 34, 45, 68 |

## III. EXPERIMENTAL DESIGN AND RESULTS

### A. Database configuration

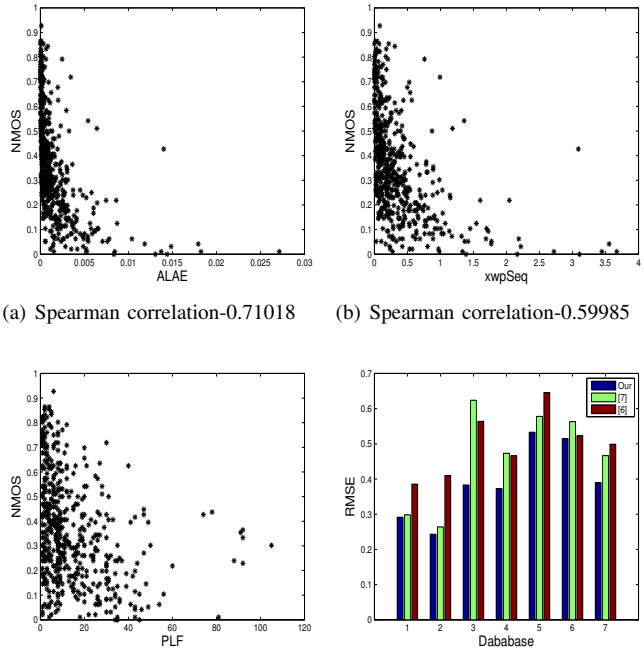
In order to develop the model, 5 training databases are built for training the coefficients in (7) and determining the function $g(.)$ in (6), and 6 validation databases for testing the model performance. Each database contains 8 video contents with 10s duration of high dimension (HD) or standard dimension (SD). The hypothetical reference circuits (HRCs) are encoded by H. 264, and packet-loss-concealment (PLC) mode is slicing and packet-loss-duration is random or burst. The configuration of training and validation databases is summarized in Table. 2. The testing environment is conformed to ITU-R BT.500 and subjective test is performed using the Absolute Category Rating with Hidden Reference method in ITU-T Rec. P.910. MOS value per HRC is the averaged rating from 24 subjects.

### B. Experimental results

Since the bit-rate, frame-rate, number of slice per frame, GOP structure and frame-type are considered in the calculation of model parameters, the trained one set of coefficient is sufficiently used in other 6 databases for cross-validation. We compared our quality prediction model with other two proposed in [3] and [4]. Similar to our method, model in [3] can estimate coding degradation using bitrate as parameter fitted by a logistic function, and loss degradation using PLF as parameter fitted by a exponential function. xwpSEQ metric proposed in [4] is applicable to slicing-type loss degradation, which is fitted by log function. The Spearman correlation of loss-related metric ALAE in our model, xwpSEQ in [4] and PLF in [3] with NMOS is shown in Fig. 3(a)-3(c), respectively. The significant outperformance of metric ALAE with xwpSEQ and PLF is evident, which indicates that ALAE metric is superior to others' and more correlated with the subjective quality. In Fig. 3(d) the Root Mean Square Error (RMSE) between the predicted and perceived quality using our model, model in [3] and model in [4] is presented. The RMSE value generated by our model is outperformed or comparative with other models in all the databases from index 1-6, and clearly better in mean value in index 7, which demonstrates its good performance.

## IV. CONCLUSION

This paper proposes a packet-layer quality assessment model for monitoring the quality of encrypted video. This




(a) Spearman correlation-0.71018 (b) Spearman correlation-0.59985

(c) Spearman correlation-0.40475 (d) RMSE

Fig. 3. Metric and model performance.

model is applicable to in-services, non-reference and non-intrusive applications, and its computational load is quite light by only using the packet-header information and avoiding access to media signals. An efficient loss-related parameter is proposed to predict the visible artifacts and perceived quality. The extraction of the parameter is based on the spatiotemporal complexity from frame-layer information. The overall model is capable of handling videos with various slice number, combined both coding and channel artifacts, either fixed or adaptive GOP length. The generality of the model is demonstrated from adequate amount of training and validation databases with various configurations. The better performance in metric correlation and RMSE comparison addresses the superiority of our model to others'.

## ACKNOWLEDGMENT

The authors would like to thank Ericsson, Deutsche Telekom, Huawei, NetScout, NTT, Technicolor, Telchemy and Yonsei University for building databases. The work described in this paper was partially supported by Young Scientist Foundation of Xi'an University of Architecture and Technology QN1304, and the National Natural Science Foundation of China under Grant 61301090.

## REFERENCES

[1] ITU-T Recommendation G. 107, "The E-model, a computational model for use in transmission planning", March,2005,

[2] ITU-T Recommendation G. 1070, "Opinion model for video-telephony applications", April, 2007,

[3] K. Yamagishi and T. Hayashi, "Parametric packet-layer model for monitoring video quality of IPTV services", *IEEE International Conference on Communications (ICC)*, 2008.

[4] M.N. Garcia and A. Raake, "Frame-layer packet-based parametric video quality model or encrypted video in IPTV services", *International Workshop on Quality of Multimedia Experience (QoMEX)*, 2011.

[5] ITU-T document, "Draft terms of reference (ToR) for P.NAMS", October, 2009. avaliable at http://www.itu.int/md/meetingdoc.asp?lang=en&parent=T09-SG12-091103-TD-GEN-0146.

[6] ITU-T document,"Draft terms of reference (ToR) for P.NBAMS", March, 2009. avaliable at http://www.itu.int/md/T09-SG12-110118-TD-GEN-0521.

[7] A. Talahashi and D. Hands and V. Barriac, "Standardization activities in the ITU for a QoE Assessment of IPTV", IEEE *Communication Magazine*, 2008, pp. 78-84.

[8] N. Liao and Z. Chen, "A Packet-Layer Video Quality Assessment Model with Spatiotemporal Complexity Estimation", EURASIP *Journal on Image and Video Processing*, 2011.