A REDUCED-REFERENCE PARAMETRIC MODEL FOR AUDIOVISUAL QUALITY OF IPTV SERVICES

Toni Mäki[1], *Dragan Kukolj*[2], *Dragana Đordević*[3], *Martín Varela*[1]

[1] VTT Technical Research Centre of Finland
Email: {Toni.Maki,Martin.Varela}@vtt.fi

[2] University of Novi Sad, Faculty of Technical Sciences
[3] RT-RK Institute, Novi Sad, Serbia
Email: {Dragan.Kukolj,Dragana.Djordjevic}@rt-rk.com

ABSTRACT

In this paper we present a parametric model for audiovisual quality estimation in IPTV and similar services. The proposed model takes advantage of signal characteristics calculated at the sender (in particular related to levels of motion in the content), but is purely parametric on the estimation (i.e. it does not require peeking into the bitstream), which makes it suitable for large-scale real-time monitoring applications. In order to obtain the model, we followed the Pseudo-Subjective Quality Assessment (PSQA) methodology, and compared different kinds of statistical estimators, namely Multilayer Perceptrons (MLP) and Random Neural Networks (RNN).

Index Terms— audiovisual quality assessment, video quality assessment, feature clustering, neural network

1. INTRODUCTION

The rise of IP-based TV services and the migration from traditional cable to them poses non-trivial challenges to service providers. Customers expect certain (high, comparable to cable service) levels of quality, which in turn makes managing the performance of the delivery networks and systems important. A key factor of this management is a good understanding of how the delivery of the stream affects its quality. Moreover, it is important to be able to react to the dynamic nature of network performance, and so the estimations need to be performed in real-time.

Estimating video quality, audio quality and audiovisual quality objectively are all well established research topics. The models used in estimation can be classified by the type of inputs used and the way they utilise the original signal in calculating QoE [1]. *Signal-based* and *bitstream models* rely on partially or fully decoded received payload, whereas *parametric models* typically use only packet header and optional side information. Full-reference models (FR) use the original signal as a reference; reduced-reference (RR) models use certain features of the original signal, and compare them to the same features extracted from the received signal. In contrast, no-reference (NR) models work independently of the original signal, either by analyzing the received signal, or by taking a parametric approach.

In this paper we propose a parametric RR model, i.e. one that uses features of the original signal not to compare to the same features of the received one, but to feed a function that will produce a quality estimate. The function is implemented by means of a Neural Network – in this work we tested and compared two families of NNs. In order to obtain the ground truths needed to train the model, we conducted a subjective audiovisual quality assessment campaign for H.264/AAC streams in IPTV-like network conditions, where the impact of several quality-affecting factors was considered. The subjective study utilised an experimental extension to the voting and evaluation procedure recommended in [2], designed to collect audiovisual, audio and video scores within a single assessment session.

The rest of the paper is structured as follows. Section 2 presents some of the main related work. Section 3 describes the subjective campaign. In Section 4 we present the results of the study and describe the model and the performance of the two neural network approaches. We conclude the paper in section 5 and provide some pointers to further work in this area.

2. RELATED WORK

FR models have been developed for audio and video QoE estimations with good success. However, FR models do not meet the real-time requirements of network measurements, since the original signal is rarely available where the estimations are needed. We therefore focus on NR or RR models.

Signal-based and Bitstream NR quality prediction models are generally more suitable for real-time network measurements and perform well in terms of accuracy. For example, the ITU-T P.563 [3] metric provides quite accurate estimates of listening quality for voice applications. Chin et al. [4] showed that a NR bitstream-based video quality metrics can be accurate. While these types of signal-based NR models are sometimes suitable to network-based quality monitoring, they have two drawbacks. Firstly, they are computationally expensive, which limits the scale of their deployment. Secondly, operators often deliver the media streams encrypted. Decryption of the payload, when feasible, adds to the computational cost of the monitoring, and in the general case, is unfeasible. Parametric models typically avoid these limitations as they usually do not consider the content signal at all. Garcia and Raake [5] demonstrated that parametric models can be tuned to specific use case very accurately with thorough understanding of the content delivery protocols and selecting the appropriate side information. The predictions of their video QoE model have excellent correlation with subjective assessment for both SD and HD content.

There are certain key aspects to consider when assessing audiovisual quality. It is known that in non-conversational services the video modality dominates over the audio modality in how they contribute to the overall audiovisual quality [2]. The A/V synchronisation also influences the audiovisual QoE. One way to calculate estimation of audiovisual quality is to combine separate estimations of audio and video quality into a single approximation. Annex C of P.911 [2] proposes the form $MOS_{av} = \alpha + \beta \cdot MOS_a \cdot MOS_v$, that maps one-way audio quality and one-way video quality into audiovisual quality. Several studies found the mapping to hold accurately. However, Garcia et al. [6] pointed out a shortcoming in this approach. The first audiovisual model presented by them is an equation that uses pre-calculated audio and video qualities; $Q_{av} = \alpha + \beta \cdot Q_a + \gamma \cdot Q_v + \zeta \cdot Q_a \cdot Q_v$. They showed that this model could not fully capture the influence of impairments (mainly audio related) to integral audiovisual quality. The second method presented in [6] uses directly lower level metrics in audiovisual quality estimation, resulting in quality predictions with better correlation.

Machine learning techniques have been successfully applied to estimating quality. Multilayer perceptrons (MLP) have been used for FR image assessment [7], [8], FR video assessment [9] and NR video assessment [10]. The PSQA [11] methodology typically uses Random Neural Networks (RNN), and has been successfully used for NR video and voice quality estimation [12], [13].

While neural networks have been often used to build video QoE models, audiovisual models based on NN are not that common. Han et al. [14] proposed a signal-based audiovisual quality model based on NN approach with moderately good results.

3. SUBJECTIVE ASSESSMENT

We conducted a subjective assessment campaign following guidelines of ITU-T recommendation P.911 [2] as closely as was feasible except for the rating mechanism. We altered the Degradation Catecory Rating (DCR) method as described in 3, in order to collect audiovisual, audio and video ratings in a single session.

Dimension	Description	Values
RES	The resolution of video sample	854x480, 1280x720, 1920x1080
MQ	The amount of movement (subjectively evaluated) of the viewed sample.	Low, Moderate, High
LR	Percentage of packets being lost during the transmission of the video sample	0.3 %, 0.6 %, 1.2 %, 2.4 %, 4.8 %
MLBS	Mean loss burst size during the transmission of the video sample	1.0, 2.0, 3.0
MQ-C	Sum of TI and SI calculated for original sample	80.9 – 135.2

Table 1: The influence factors studied

The campaign was planned to consider the impact of resolution (RES), movement quantity (MQ), loss rate (LR), mean loss burst size (MLBS), and error concealment (EC). Due to an error in the instrumentation, the error concealment was not varied, and hence cannot be considered as planned. However, this error led to an interesting discovery regarding the impact of loss rates, as discussed in section 4. The MQ levels were chosen to be low, medium and high, and content selected accordingly. In order to train the models, the MQ levels were substituted by a simple additive metric based on actual content-based calculations of TI (scene cut Temporal Perceptual Information and SI (Spatial Perceptual Information [15]). The influence factors, plus the additional calculated Movement Quantity (MQ-C) are presented with their respective value ranges are summarized in Table 1.

The original audiovisual material was downloaded from The Consumer Digital Video Library [16]. The selected samples are listed in Table 2. Four video samples were selected to represent each level of MQ. The initial classification into levels of MQ was based on the visual assessment of the videos, on the meta-information related to original video samples and on the advice of calculated TI. The "NTIA Front End (Part of a Longer Talk)" sample was used in two sequences, displaying different sections of the original sample. The frame rates of the original video samples were either 25 fps or 30 fps.

The original video samples were edited as per the P.911 guidelines and encoded using H.264. The samples were encoded with 2-pass method (baseline profile) into different bitrates depending on their resolution (6 Mbps for 1920x1080, 3 Mbps for 1280x720 and 1 Mbps for 854x480) using the x264 encoder. The slices of H.264 encoded content were restricted to 1460 bytes in order to fit a single slice in a single UDP packet (this was taken into account also in the packetisation). Intra frame interval was configured to a maximum of 1 s. The audio was encoded as AAC into two different bitrates (96 kbps for 1920x1080 and 1280x720 resolutions and

Original sample	Description
NTIA outdoor mall with tulips (3e)	A view of a pedestrian mall.
NTIA Front End (Part of a Longer Talk)	Part of a description of a device.
NTIA Snow Mountain	A sequence of mountains and trees.
NTIA snowy day in the city (1e)	Three scenes with snowfall.
NTIA Jump Rope	A man jumping a rope.
NTIA Elephant Crane	An elephant crane playing on a stick.
NTIA Highway Cuts	Several views of cars driving.
PSCR Score Narrow	A football (American) player after scoring a touchdown.
PSCR Touchdown Day	A football (American) player scoring a touchdown.
NTIA Cheetah	A cheetah walking in a cage.
NTIA Red Kayak	A man rowing a red kayak

Table 2: Original video samples

64 kbps for 854x480 resolution).

The participants consisted of 24 VTT employees (7 female, 17 male). Ten persons had prior experience with multimedia quality assessment, 4 were multimedia experts without prior experience of quality assessments, 7 were technical people (non-multimedia, non-quality assessment) and 3 subjects were considered by themselves non-technical. Twenty of the 24 subjects were native Finnish speakers. The subjects' age ranged between 24 and 46 years old, and they all had normal or corrected vision and hearing.

The test sequences were prepared prior to the assessment by recording RTP-based video streams transmitted over an emulated network. The videos were streamed and recorded with the VLC media player[1] from an instance of the Darwin Streaming Server (DSS)[2]. A network emulator — Netem was deployed in order to introduce the desired LR and MLBS for each test condition. The regular Netem shipped as part of the Linux kernel implements only simplistic loss models, and cannot guarantee to introduce a precise loss rate over short intervals (in long intervals the realized loss rate approaches the expected value). Therefore the Netem CLG fork[3] was used, which allows, among other things, using pre-calculated loss traces. The predefined loss patterns for each test condition were in turn generated with an in-house tool implementing a simplified Gilbert-Elliott model[4].

Given the large size of the parameter space, and following the PSQA approach, a smart sampling scheme was deployed to reduce the number of conditions to be tested. With the smart sampling we enforced some conditions in the border areas of the parameter space, and emphasised the presence of the most realistic conditions of the space during otherwise random sampling. A tool was developed to perform the sampling, allowing to weigh and enforce different factors separately[5]. After sampling, a total of 125 conditions were considered.

The viewing and listening conditions specified in Section 7.1 of P.911 [2] were followed as much as was feasible. The presentation of the sequences and voting the difference was done according to the DCR method described in P.911 with a modified voting procedure in which after each sequence the subjects were presented with three questions instead of one. In the first part of the voting (lasting 7 seconds) subjects were asked to rate the difference in audiovisual quality of the pair of sequences. In the second part of the voting (lasting 10 seconds), subjects were asked to rate separately the difference in audio quality and the difference in video quality. The altered viewing and voting procedure is presented in Figure 1. The five-level impairment scale from Table 3 in P.911 (Very Annoying, Annoying, Slightly Annoying, Perceptible but not annoying, Imperceptible) was used in voting. The order of the rendered sequences was randomly drawn before the assessments, but was same for all the subjects. Instructions were given in paper format in both Finnish and English. The instructions were based on Section II.2 of Appendix II of P.911. Subjects were allowed to ask questions. After reading the instructions, subjects performed a training session. The full set of sequences consisted of 125 sequences. There was a pause after 70th sequence and total duration of the assessment session was about 1.5h (depending on the length of the pause each subject chose to have between both parts).

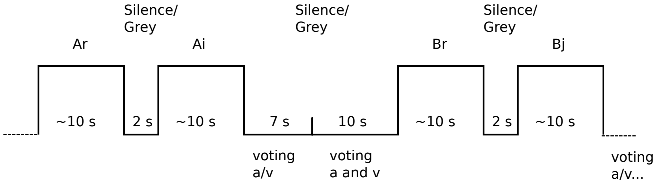


Fig. 1: Modified stimulus presentation in the DCR method

4. RESULTS AND MODELLING

4.1. Observations on subjective data

The voting behaviour of the subjects was analyzed with SOS [17] (Standard deviation of the scores). 'SOS a' values found for the audiovisual, audio and video assessments were 0.1154, 0.167779 and 0.126404, respectively. These values are very close to those calculated for H.264/AVC quality assessment in [17], and they suggest that the subjects were able to assess the quality rather consistently. The subjects were least certain when giving votes of audio quality. The correlations of the audio quality, video quality and interaction of these two with the audiovisual quality were calculated. As it was expected, video quality has higher correlation (0.95)

[1] http://videolan.org

[2] http://dss.macosforge.org/

[3] https://netgroup.uniroma2.it/twiki/bin/view/Main/NetemCLG

[4] https://github.com/mvarela/Gilbert-Loss-Trace-Generator

[5] This tool is freely available for research purposes, please contact the authors for further information.

IF	$R_s(av)$	$R_p(av)$	$R_s(v)$	$R_p(v)$	$R_s(a)$	$R_p(a)$
RES	-0.219	-0.231	-0.312	-0.313	-0.030	-0.030
MQ	-0.159	-0.163	-0.267	-0.259	-0.002	0.034
MLBS	0.062	0.063	0.038	0.041	0.097	0.060
LR	*-0.818*	*-0.763*	*-0.797*	*-0.731*	*-0.726*	*-0.687*
MQ-C	0.237	0.245	0.274	0.281	0.104	0.088

Table 3: Spearman and Pearson correlation of input data to actual MOS for audio-visual, video and audio modalities

with audiovisual quality than audio quality (0.82). The interaction of audio and video quality has the highest correlation (0.98) with audiovisual quality than either of the individual qualities.

The significant main effects of the input variables on observed MOS values were investigated as part of the MLP training process. Table 3 summarises the Spearman and Pearson correlations of the input variables and MOS for the different modalities. LR has the largest effect on all the modalities. While audio quality is mainly dependent on LR, audiovisual and video quality are also affected by MQ-C and RES. The significance of these effects was investigated with a set of one-way ANOVAs and they were found significant (P-values below 0.05). The MLBS did not have a significant effect. This may be explained partly by the lack of loss concealment due to the instrumentation bug (cf. similar effect for MLBS and FEC in voice LQ [11]), and partly by the very small presence of B-frames in the encoded video samples. Normally an isolated packet loss has a larger probability of dropping a B-frame than other frame types. Given the low amount of B-frames in the GOP structure used, the effect of isolated losses on the perceived quality is likely increased.

An observation worth noting was made regarding the conditions in which the configured parameters were the same, but the actual loss patterns were different (LR and MLBS being equal). In some cases, this led to differences of up to 1 MOS point in the subjective assessment, and suggests that models incorporating bitstream-level knowledge (i.e. knowing which frame types were affected) will perform better than those using only packet-level information, and that this effect should be taken into account in further subjective studies.

4.2. A Parametric RR Model for Audiovisual Quality

As seen from the parameter correlations with subjective scores, the calculated motion quantity parameter has a stronger impact than the subjectively assigned *a-priori*. Normally, this information is not available in real-time, as the reference signal is not available, and the received one might be degraded, so the calculation is not necessarily accurate. Moreover, it adds to the computational complexity of the assessment. In some scenarios, notably those such as IPTV, where the video is basically multi-casted, it would be relatively simple to pre-calculate this at the source, and then transmit the information as accompanying meta-data to be used by a parametric quality model on the receiver side. The actual implementation of such a protocol is beyond the scope of this paper.

	LR	LR, RES	LR, RES, MQ-C	All IFs
$R_s(av)$	0.813	0.880	0.920	0.961
$R_p(av)$	0.850	0.923	0.954	0.978
RMSE	0.415	0.294	0.235	0.150
σ_{RMSE}	0.009	0.011	0.033	0.027

Table 4: Performances of MLP-based estimators for audio-visual streams using different sets of influence factors as input

We therefore propose a PSQA-based model that takes this information into account, as well as the other parameters listed in Table 1.

4.2.1. Difference in Neural Networks Performance

In [11], it was shown that RNNs performed better than traditional ANNs for PSQA-based voice quality estimations, and that the estimators' performance is somewhat dependent on the NN topology used. In this work we implemented the PSQA models with both MLPs and RNNs. For RNNs, several three-layer topologies were tested, and for both NN types, the resulting estimators were evaluated by means of 10-fold cross-validation. Interestingly, in this application, the MLP performed significantly better than the RNNs.

The MLP was configured to have a single hidden layer, and the number of input neurons equaled the number of the input features. The number of the hidden neurons was defined by formula $2n + 1$, where n is the number of inputs. The tangent-hyperbolic function was chosen as the activation function of the hidden nodes, while linear function was chosen for the output neuron. Training of MLP was done by Levenberg-Marquardt [18] [19] optimization back-propagation algorithm.

Several MLP based estimators were built using different input parameter sets. Table 4 summarises the performances of MLP based audiovisual models trained and tested with different input parameter combinations. The accuracy of the predictions calculated for the test sequences improves with respect to the number of the influence factors considered by the models.

The performances of MLP based audiovisual, video and audio QoE estimators using all available information are given in Table 5. The figures are calculated comparing the predicted MOS and actual MOS of the test sets (each predicted MOS in this calculation is an average of 10 predicted MOS values got from ten-fold cross-validations). It can be seen that both audio-visual and video QoE models can estimate the MOS accurately. The audio QoE model on the other hand, does not perform that well. The audiovisual QoE estimation accuracy related to a single cross-validation run is illustrated in Figure 2a.

The RNNs were trained using all influence factors as input parameters. A set of different neural network architec-

Modality	R_s	R_p	RMSE	σ_{RMSE}
Audiovisual	0.961	0.978	0.150	0.027
Video	0.963	0.980	0.144	0.034
Audio	0.657	0.654	0.691	0.038

Table 5: Performances of MLP-based MOS estimator for audio-visual, video and audio data streams

Modality	R_s	R_p	RMSE	σ_{RMSE}
Audiovisual	0.802	0.846	0.485	0.096
Video	0.862	0.861	0.486	0.113
Audio	0.642	0.703	0.702	0.153

Table 6: Performances of RNN-based MOS estimators for audio-visual, video and audio data streams

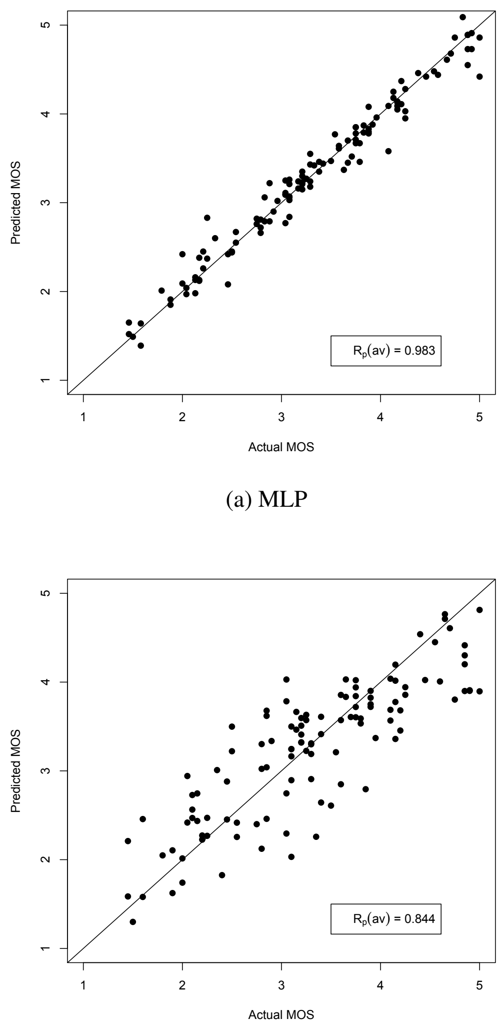
tures was iterated in order to identify the best fitting architecture (i.e. the optimal number of hidden neurons). The architectures investigated had minimum of 9 and maximum of 15 hidden neurons. The number of input neurons was the number of IFs. The best architecture was then selected based on average correlations and RMSE, reflecting the best observed behaviour. Table 6 summarises the performance of the best fitted models. Figure 2b illustrates the accuracy of the audio-visual predictions of the selected neural network architecture (MOS values are from the test sets of the concerned ten-fold cross-validation). Architectures with 9 hidden neurons, 14 hidden neurons and 12 hidden neurons were identified as the best option for audio-visual, video and audio quality, respectively.

5. CONCLUSIONS AND FURTHER WORK

We have presented a reduced-reference parametric model for audiovisual quality estimation, and studied its performance. The model follows the PSQA methodology, and performs at its best when implemented with a multilayer perceptron — MLP.

We trained the model with subjective assessment data for an IPTV-like scenario, which is the target application for it. The estimation accuracy of MLP-based implementation shows that with careful optimisation, the neural networks can be used to build high-precision audiovisual models. The difference in the accuracy of the two NN models underlines the need of the careful engineering in fine-tuning the models in order to avoid over- and under-fitting.

While signal-based models and bitstream models usually provide more accurate estimations than parametric models that exploit only packet header information, we showed that by adding a small amount of information about the original signal, the performance of these packet-level models can be very good for certain uses. The correlations obtained with the model are most likely good enough for many real world monitoring solutions needed for network management purposes, and they scale extremely well, since their computational cost is negligible, and the reference information needs only be calculated once at the source (and possibly offline).



(a) MLP

(b) RNN

Fig. 2: Actual MOS vs predicted MOS.

From the subjective assessment results, there is a lesson to learn from the altered viewing and voting procedure. The SOS analysis and the low correlations achieved in the neural network cross-validations for the audio modality imply that the used method of acquiring audio quality should be improved, better taking into account the dominance of the video modality.

This study has shown the viability and practicality of our modeling approaches in a specific application setting. In order to widen the range of applicability the correct side information related to other scenarios should be identified. This requires making studies that take into account different error concealment strategies, the task and context of the audience and longer term effects, just to name a few.

6. ACKNOWLEDGEMENTS

This work was supported in part by Ministry of Science and Technological Development of Republic of Serbia, under 32034 project and Secretary of Science and Technology Development of Vojvodina Province under Grant 114-451-2434/2011-01 and Tekes, the Finnish Funding Agency for Technology and Innovation, under the CELTIC IPNQSIS project.

7. REFERENCES

[1] A. Raake, J. Gustafsson, S. Argyropoulos, M. Garcia, D. Lindegren, G. Heikkila, M. Pettersson, P. List, and B. Feiten, "IP-Based mobile and fixed network audiovisual media services," *IEEE Signal Processing Magazine*, vol. 28, no. 6, pp. 68–79, Nov. 2011.

[2] ITU-T Recommendation P.911, "Subjective audiovisual quality assessment methods for multimedia applications," 2010.

[3] ITU-T Recommendation P.563, "Single-ended method for objective speech quality assessment in narrow-band telephony applications," 2004.

[4] M. Chin, T. Brandao, and M.P. Queluz, "Bitstream-based quality metric for packetized transmission of h.264 encoded video," in *2012 19th International Conference on Systems, Signals and Image Processing (IWSSIP)*, Apr. 2012, pp. 312–315.

[5] M.N. Garcia and A. Raake, "Parametric packet-layer video quality model for IPTV," in *2010 10th International Conference on Information Sciences Signal Processing and their Applications (ISSPA)*, May 2010, pp. 349–352.

[6] M. N. Garcia, R. Schleicher, and A. Raake, "Impairment-factor-based audiovisual quality model for IPTV: influence of video resolution, degradation type, and content type," *EURASIP Journal on Image and Video Processing*, vol. 2011, no. 1, pp. 1–14, Mar. 2011.

[7] A. Bouzerdoum, A. Havstad, and A. Beghdadi, "Image quality assessment using a neural network approach," in *Signal Processing and Information Technology, 2004. Proceedings of the Fourth IEEE International Symposium on*, dec. 2004, pp. 330–333.

[8] S. Kaya, M. Milanova, J. Talburt, B. Tsou, and M. Altynova, "Subjective image quality prediction based on neural network," in *Lecture Notes in Proceedings of 16th Int. Conf. on Information Quality ICIQ*, 2011, pp. 259–266.

[9] H. El Khattabi, A. Tamtaoui, and D. Aboutajdine, "Video quality assessment measure with a neural network," *International Journal of Computer and Information Engineering*, vol. 4, no. 3, pp. 167–171, June 2010.

[10] D. Kukolj, M. Pokrić, V. Zlokolica, J Filipović, and M. Temerinac, "No-reference video quality assessment design framework based on modular neural networks," in *Lecture Notes in Computer Science*, vol. 6352, pp. 569–574. Springer-Verlag, Berlin, Heidelberg, 2010.

[11] Martín Varela, *Pseudo–Subjective Quality Assessment of Multimedia Streams and its Applications in Control*, Ph.D. thesis, INRIA/IRISA, univ. Rennes I, Rennes, France, Nov. 2005.

[12] S. Mohamed and G. Rubino, "A study of real-time packet video quality using random neural networks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 12, no. 12, pp. 1071–1083, Dec. 2002.

[13] Ana Couto da Silva, Martín Varela, Edmundo de Souza e Silva, Rosa Leão, and Gerardo Rubino, "Quality assessment of interactive real time voice applications," *Computer Networks*, vol. 52, pp. 1179–1192, Apr. 2008.

[14] Xinlu Han, Yaodu Wei, and Xiang Xie, "An audio-visual objective quality model based on BP neutral network," in *2011 International Conference on Multimedia Technology (ICMT)*, July 2011, pp. 5277–5280.

[15] ITU-T Recommendation P.910, "Subjective video quality assessment methods for multimedia applications," 2008.

[16] Intel Corporation, National Telecommunications and Information Administration's Institute for Telecommunication Sciences and University of California at Santa Barbara, "The consumer digital video library," .

[17] T. Hoßfeld, R. Schatz, and S. Egger, "SOS: the MOS is not enough!," in *2011 Third International Workshop on Quality of Multimedia Experience (QoMEX)*, Sept. 2011, pp. 131–136.

[18] K. Levenberg, "A method for the solution of certain problems in least squares," *Quart. Applied Math.*, vol. 2, pp. 164–168, 1944.

[19] Donald W. Marquardt, "An Algorithm for Least-Squares Estimation of Nonlinear Parameters," *SIAM Journal on Applied Mathematics*, vol. 11, no. 2, pp. 431–441, 1963.