ITUPublications

International Telecommunication Union
Standardization Sector

ITU-T Technical Report

(01/2025)

PSTR-CMVTQS1

Alternative computational model used as a quality monitor to assess videotelephony services based on machine learning

Technical Report ITU-T PSTR-CMVTQS1

Alternative computational model used as a quality monitor to assess videotelephony services based on machine learning

Summary

This Technical Report describes a computational model based on machine learning (ML) techniques that can be used as a quality monitor to assess videotelephony services. It has been developed within the framework of ITU-T Study Group 12's "G.CMVTQS" work item project and is an alternative approach to the model described in Recommendation ITU-T P.940. Clauses of this report address the presentation of ML techniques used, the architecture of the model, the various modules of the model and their formulae, as well as the performance of the model on the databases used for the validation of Recommendation ITU-T P.940.

Keywords

Machine learning, QoE, videotelephony.

Note

This is an informative ITU-T publication. Mandatory provisions, such as those found in ITU-T Recommendations, are outside the scope of this publication. This publication should only be referenced bibliographically in ITU-T Recommendations.

Editor:	Zhenzhong Chen	Wuhan University
	Yaosi Hu	Wuhan University
	Mengying Liu	China Mobile Communication Co. Ltd.
	Lei Yang	China Mobile Communication Co. Ltd.

© ITU 2025

All rights reserved. No part of this publication may be reproduced, by any means whatsoever, without the prior written permission of ITU.

Table of Contents

Technical Report ITU-T PSTR-CMVTQS1

Alternative computational model used as a quality monitor to assess videotelephony services based on machine learning

1 Machine learning model

The machine learning model described in this Technical Report employs the random forest method to build the model for quality of experience (QoE) assessment of videotelephony services. The random forest (RF) algorithm [b-Breiman], learns a set of decision trees on a randomly sampled subset of the features of training data with replacement (bootstrapping). The scikit-learn toolbox's RF regression algorithm was used as the regressor [b-Pedregosa]. A total of 100 trees, or estimators, were learnt using this bootstrapping method. Every tree received all of the features and training sample set for training. Each of the 100 trees splits its subset of data until all leaves are pure or the maximum depth is reached. Each tree receives eight randomly sampled features for learning. Gini impurity was used as the measure of node split quality. The operational principles of random forest are illustrated in Figure 1, in which θ_i represents the feature vector, h_i represents the decision tree, and L_i represents the result.

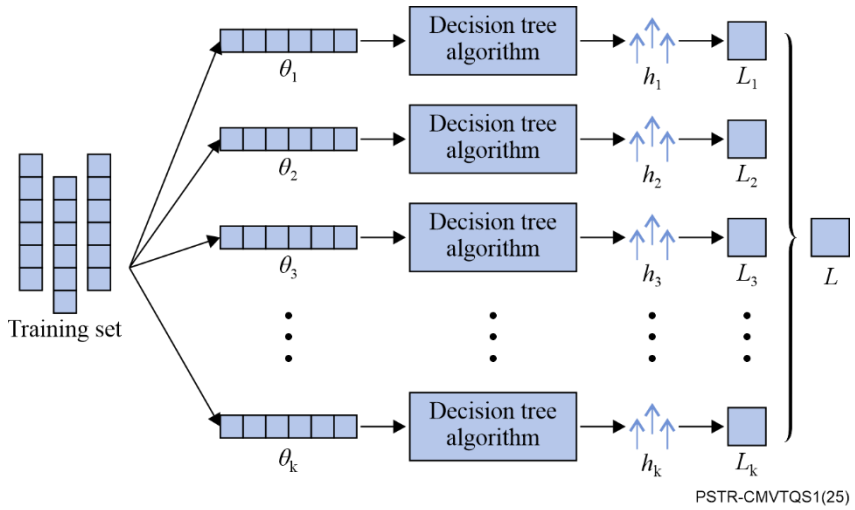

Figure 1 – Random Forest algorithms

The unified hyperparameter settings for each module are shown in Table 1. These hyperparameters were chosen empirically to maximize performance while mitigating overfitting.

Table 1 – ML algorithm parameterization and settings

ML algorithm	Parameterization	Settings
RF regressor	num_trees	100
	min_samples_per_leaf	1
	feature_subsampling_fraction	1

The database was randomly sampled into five cross-validation (CV) sets of 80% training and 20% testing. For each CV split, all parameters in the feature space were included. Random sampling ensured that we tested the model on unknown data. The RF algorithm was trained on the training split and tested on the testing split.

Below is the pseudocode that illustrates the training and evaluation process of the model:

```
from sklearn.ensemble import RandomForestRegressor
input = correspon
label = MOS
num_trees = 100
min_samples_per_leaf = 1
feature_subsampling_fraction = 1.0
random_state = 42
X_train, X_test, y_train, y_test = train_test_split(input, label, test_size = 0.2, random_state = random_state)
M = RandomForestRegressor(n_estimators = num_trees, min_samples_leaf = min_samples_per_leaf, max_features = feature_subsampling_fraction, random_state = random_state )
M.fit(X_train, y_train)
Q = M.predict(X_test)
Evaluate(Q, y_test)
```

2 Model description

2.1 Video quality assessing block

To estimate video quality during video calls, the quality of a video segment in the one-way communication scenario, which is simulated by the non-interactive audiovisual material subjective tests described in Appendix II of [b-ITU-T P.940], is defined as basic video quality (Q_{vs}). This metric is predicted using a random forest model that is trained on a set of features dependent on the terminal factors and video quality factors defined in clause 6.1 of [b-ITU-T P.940]. The input features are drawn from the following parameter space:

$$P = \{Br_v, Fr_v, R_h, R_w, D_s\} \tag{1}$$

where, Br_v is the video bit rate, Fr_v is the video frame rate, R_h is the video resolution height, R_w is the video resolution width, and D_s is the screen size (inch). Each element of the parameter space P corresponds to a specific combination of these parameters. Q_{vs} for a given set of parameters can then be expressed as:

$$Q_{vs} = M_{dt}(P) = M_{dt}(Br_v, Fr_v, R_h, R_w, D_s) \tag{2}$$

where, $M_{dt}(.)$represents the function learned by the random forest model during training for the particular device type (D_t).

NOTE – The function of Q_{vs} were developed for three different types of terminal devices, namely mobile, personal computer (PC), and television (TV).

In scenarios where the terminal factor D_s is unknown, the model predicts the video quality based on the reduced feature set $P = \{Br_v, Fr_v, R_h, R_w\}$, and Q_{vs} is estimated by:

$$Q_{vs} = M_{dt}(Br_v, Fr_v, R_h, R_w) \tag{3}$$

In the context of two-way communication which is emulated by the interactive conversational subjective tests described in Appendix II of [b-ITU-T P.940], network transmission impairment factors, such as packet loss, can significantly impact video quality according to [b-ITU-T G.1070]. To account for these effects, the video quality is defined by constructing a parameter space that combines the estimated basic video quality (Q_{vs}) with network transmission impairment factors including video packet loss rate (Plr_v) and video interarrival jitter (jit_v).

The video quality Q_v affected by network transmission is thus modelled as follows:

$$Q_v = M_v(Q_{vs}, Plr_v, jit_v) \tag{4}$$

where, $M_v(.)$ is a trained RF model. Q_v is constrained by the range of MOS between 1 and 5.

2.2 Audiovisual quality assessing block

The audiovisual quality score (Q_{av}) is calculated by integrating the audio quality (Q_a) and video quality (Q_v) using a trained RF regressor ($M_{av}(.)$). The fusion of these two modalities considers both the quality of audio stream and video stream. The audiovisual quality score(Q_{av}) on a scale of 1-5 is computed as:

$$Q_{av} = M_{av}(Q_a, Q_v) \tag{5}$$

where, Q_a represents the estimated audio quality, and this ML model for quality assessment of videotelephony services adopts the Fullband version of E-model as defined in [b-ITU-T G.107.2] as a provisional method.

2.3 Audiovisual interaction delay assessing block

The definition of the audio-visual delay impairment factor in [b-ITU-T G.1070] shows a linear relationship between the absolute audio-visual delay and the combination of video delay (T_v) and audio delay (T_a). Therefore, the modelling of audio-visual interaction delay quality (Q_{delay}) is defined as:

$$Q_{delay} = M_{delay}(T_v + T_a, T_v, T_a) \tag{6}$$

where, $M_{delay}(.)$ is a trained RF regression function. The value of Q_{delay} is bounded between 1 and 5.

2.4 Audiovisual media synchronization assessing block

The definition of the audio-visual delay impairment factor in [b-ITU-T G.1070] shows a linear relationship between the audio-visual media synchronization and the difference between audio delay (T_a) and video delay (T_v). Therefore, the modelling of audio-visual media synchronization quality (Q_{sync}) is defined as:

$$Q_{sync} = M_{sync}(T_v - T_a, T_v, T_a) \tag{7}$$

where, $M_{sync}(.)$ is a trained RF model. The value of Q_{sync} is bounded between 1 and 5.

2.5 Videotelephony quality assessing block

Based on the model layout defined in clause 6 of [b-ITU-T P.940], the final videotelephony quality (Q_{vt}) on a scale of 1-5 is defined as:

$$Q_{vt} = M_{vt}(Q_{av}, Q_{delay}, Q_{sync}) \tag{8}$$

where, $M_{vt}(.)$ is a trained RF model. Q_{av} reflects the multimedia quality of audio and video during a video call, influenced by network transmission, coding distortion and terminal display. Q_{delay} and Q_{sync} represent the estimated audio-visual interaction delay quality and audio-visual media synchronization quality, respectively. Q_{delay} and Q_{sync} combine audiovisual quality and network transmission delay parameters to predict user experiences of delay and synchronization in constrained transmission channels, which are crucial for assessing the quality of bidirectional call experiences.

3 Performance figures

The performance of the model on the databases described in Appendix II of [b-ITU-T P.940] is summarized in Table 2 below.

Table 2 – Performance of ML algorithm

Quality evaluation perspective	PLCC	RMSE
Video quality (with both passive and active test data)	0.980	0.185
Video quality (with active test data)	0.965	0.306
Audio quality	0.915	0.785
Audiovisual interaction delay quality	0.977	0.235
Audiovisual media synchronization quality	0.994	0.128
Final videotelephony quality	0.977	0.228

NOTE 1 – These performance figures were calculated using five-fold cross-validation.

NOTE 2 – The performance figures of video quality assessing block were calculated using both passive and active test data, and only active test data, separately.

NOTE 3 – The impact of delay is considered separately in the audiovisual interaction delay assessing block and the audiovisual media synchronization assessing block. Therefore, the validation of video quality assessing block is conducted based on subjective test data with video delay smaller than 100 ms, and the validation of audio quality assessing block is conducted based on subjective test data with audio delay smaller than 100 ms.

Bibliography

[b-ITU-T G.107.2] Recommendation ITU-T G.107.2 (2023), *Fullband E-model.*

[b-ITU-T G.1070] Recommendation ITU-T G.1070 (2018), *Opinion model for video-telephony applications.*

[b-ITU-T P.940] Recommendation ITU-T P.940 (2025), *Computational model used for the monitoring and quality assessment of videotelephony services.*

[b-Breiman] Breiman, L. (2001), *Random forests.*

[b-Pedregosa] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O. and Duchesnay, É. (2011), *Scikit-learn: Machine learning in Python*. Journal of machine Learning research.