International Telecommunication Union

ITU-T

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

Series P

Supplement 28

(09/2020)

SERIES P: TELEPHONE TRANSMISSION QUALITY, TELEPHONE INSTALLATIONS, LOCAL LINE NETWORKS

Considerations for the development of new QoS and QoE related objective models to be embedded in Recommendations prepared by ITU-T Study Group 12

ITU-T P-series Recommendations – Supplement 28

ITU-T P-SERIES RECOMMENDATIONS

TELEPHONE TRANSMISSION QUALITY, TELEPHONE INSTALLATIONS, LOCAL LINE NETWORKS

Vocabulary and effects of transmission parameters on customer opinion of transmission quality	P.10–P.19
Voice terminal characteristics	P.30–P.39
Reference systems	P.40–P.49
Objective measuring apparatus	P.50–P.59
Objective electro-acoustical measurements	P.60–P.69
Measurements related to speech loudness	P.70–P.79
Methods for objective and subjective assessment of speech quality	P.80–P.89
Voice terminal characteristics	P.300–P.399
Objective measuring apparatus	P.500–P.599
Measurements related to speech loudness	P.700–P.709
Methods for objective and subjective assessment of speech and video quality	P.800–P.899
Audiovisual quality in multimedia services	P.900–P.999
Transmission performance and QoS aspects of IP end-points	P.1000–P.1099
Communications involving vehicles	P.1100–P.1199
Models and tools for quality assessment of streamed media	P.1200–P.1299
Telemeeting assessment	P.1300–P.1399
Statistical analysis, evaluation and reporting guidelines of quality measurements	P.1400–P.1499
Methods for objective and subjective assessment of quality of services other than speech and video	P.1500–P.1599

For further details, please refer to the list of ITU-T Recommendations.

Supplement 28 to ITU-T P-series Recommendations

Considerations for the development of new QoS and QoE related objective models to be embedded in Recommendations prepared by ITU-T Study Group 12

Summary

Supplement 28 to ITU-T P-series Recommendations provides guidelines for Recommendations that describe or specify tools for the objective estimation of dimensions of quality of service (QoS) and quality of experience (QoE) with quality models, and which are planned to be approved by ITU-T Study Group 12 (SG12).

History

Edition	Recommendation	Approval	Study Group	Unique ID*
1.0	ITU-T P Suppl. 28	2020-09-11	12	11.1002/1000/14495

Keywords

Artificial intelligence, continuous learning, guidelines, machine learning, models, QoE measurement, recommendation, requirement specification.

* To access the Recommendation, type the URL http://handle.itu.int/ in the address field of your web browser, followed by the Recommendation's unique ID. For example, http://handle.itu.int/11.1002/1000/11830-en.

FOREWORD

The International Telecommunication Union (ITU) is the United Nations specialized agency in the field of telecommunications, information and communication technologies (ICTs). The ITU Telecommunication Standardization Sector (ITU-T) is a permanent organ of ITU. ITU-T is responsible for studying technical, operating and tariff questions and issuing Recommendations on them with a view to standardizing telecommunications on a worldwide basis.

The World Telecommunication Standardization Assembly (WTSA), which meets every four years, establishes the topics for study by the ITU-T study groups which, in turn, produce Recommendations on these topics.

The approval of ITU-T Recommendations is covered by the procedure laid down in WTSA Resolution 1.

In some areas of information technology which fall within ITU-T's purview, the necessary standards are prepared on a collaborative basis with ISO and IEC.

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.

INTELLECTUAL PROPERTY RIGHTS

ITU draws attention to the possibility that the practice or implementation of this publication may involve the use of a claimed Intellectual Property Right. ITU takes no position concerning the evidence, validity or applicability of claimed Intellectual Property Rights, whether asserted by ITU members or others outside of the publication development process.

As of the date of approval of this publication, ITU had not received notice of intellectual property, protected by patents, which may be required to implement this publication. However, implementers are cautioned that this may not represent the latest information and are therefore strongly urged to consult the TSB patent database at http://www.itu.int/ITU-T/ipr/.

Table of Contents

Introduction

With the widespread implementation of machine learning (ML) and artificial intelligence (AI), these technologies are more intensely used in conjunction with the development of objective models for the estimation of dimensions of quality of service (QoS) or quality of experience (QoE) than this was the case in the past. As will be shown in this Supplement, ML and AI based solutions for measurement applications need special attention. Some of the resulting consequences however are valid for traditional measurement techniques as well. Therefore, this Supplement provides a set of very general guidelines for ITU-T Recommendations related to the measurement of QoE and other quality related factors in order to maintain the high quality of such Recommendations.

This Supplement targets mainly, but not exclusively, future new or revised Recommendations to be consented under the ITU-T P.5x, P.7x, P.8x, P.5xx, P.70x, P.8xx, P.9xx, P.12xx, P.13xx and P.15xx series.

Supplement 28 to ITU-T P-series Recommendations

Considerations for the development of new QoS and QoE related objective models to be embedded in Recommendations prepared by ITU-T Study Group 12

1 Scope

This Supplement provides guidelines which are to be followed before moving to ITU-T SG12 Consent of ITU-T Recommendations related to the estimation of perceptual quality factors. The majority of the guidelines can also be applied to Recommendations targeting the measurement of other quality factors, for which the ground truth is not determined by human perception.

2. References

None.

3 Definitions

3.1 Terms defined elsewhere

This Supplement uses the terms defined in [ITU-T P.10].

3.2 Terms defined in this Supplement

None.

4 Abbreviations and acronyms

This Supplement uses the following abbreviations and acronyms:

AI	Artificial Intelligence
ANN	Artificial Neural Network
LSTM	Long Short Term Memory
ML	Machine Learning
MOS	Mean Opinion Score
QoE	Quality of Experience
QoS	Quality of Service

5 Conventions

Quality factors

In this Supplement, quality factors are understood to mean deterministic factors which define QoE by themselves or in combination with other QoE influencing factors as defined in clause 6.210 of [b-ITU-T P.10]. Examples of quality factors addressed in this Supplement are speech or video quality. Other factors, e.g., delay or download speed, if estimated based on other measurements instead of direct measurement, are also in the scope of this Supplement (their direct measurements would not require a measurement model and is therefore not the focus of this Supplement). QoE influencing factors related to the context of use, that is, users' expectations, cultural background, psychological profiles or emotional states of the user, are outside of the scope of this Supplement.

Artificial intelligence

In computer science, AI, sometimes called machine intelligence, is intelligence demonstrated by machines, in contrast to the natural intelligence displayed by humans.

Machine learning based model

Machine learning is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions, relying on patterns and inference instead. It is seen as a subset of artificial intelligence. Machine learning algorithms build a mathematical model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to perform the task [b-ML].

Algorithmic model

A model, whose structure is defined by fixed mathematical equations or decision rules and the task is hard defined by instructions, as opposed to models based on machine learning methods, where the resulting structure of the task is defined by the learned coefficients.

Independent validation/verification

Independent validation shall mean that a model's accuracy and suitability has been validated by another party. Alternatively, it is possible for a model proponent to validate its own model if the process is supervised by another third party, or the data used for the validation were not available, until the model was frozen.

Learning, validation and test data set

In the ML community, these terms are typically used in a way that is different from the usage in SG12. To facilitate the understanding for the outside reader, the ML terminology is used in this Supplement, which means: During the training of a model the *learning* data set is used. After several iterations, the result is verified by comparison to a *validation* data set. At the end, after the training phase, the model performance is verified using a separate *test* data set and no further training is performed.

6 Quality criteria for ITU-T Recommendations related to the measurement of quality factors

Nowadays, systems are more or less automatically tuned to improve the end user's QoE. In order to tune systems for optimal QoE, it is essential that quality factors can be determined in a cost effective, reliable and reproducible way.

Currently, this can be performed by using measurement methods which had been trained to simulate the human perception of speech or video quality, for example. Human perception is seen as a fairly stable property and as such, it varies gradually, i.e., over several years, in cases for example, when users adapt to new services or ways to use services (e.g., viewing or listening habits). Such measurement methods have been thoroughly studied, frozen and standardized. They form an important anchor of QoE-based network and system optimization.

When a Study Group standardizes a method, common sense and thus reasonable criteria are applied to assess if the method is good enough to be standardized. When it comes to the context of modelling quality factors, these criteria have evolved over the past years, but they are not documented and they differ slightly between ITU-T study groups and the Recommendations developed accordingly. Since QoE measurement is so important today and since the increasing number of more and more diverse use cases leads to a significant increase in the number of Recommendations that are related to the measurement of quality factors, it has demonstrated that formal rules should be established, which should be applied to all new Recommendations in that field and can be considered when revising existing Recommendations as far as this Supplement is relevant to them.

6.1 Classification of measurement related Recommendations

The existing ITU-T Recommendations in the scope of this Supplement can be classified into Recommendations describing one or more specified models or Recommendations forming a requirement specification with performance criteria to be fulfilled by models. The latter are often also referred to as framework Recommendations, but since this term is used with another meaning in other types of Recommendations, the following clauses refers to them as requirement specifications. A third potential type of Recommendation consists essentially of a description on how to develop a measurement model (e.g., based on machine learning methods). Clauses 6.1.1 to 6.1.3 describe these classes of Recommendations in more detail. It is expected that future ITU-T Recommendations in the scope of this Supplement can all be classified in one of these three categories.

6.1.1 Model description

Recommendations describing one or multiple models contain an exact definition of the model itself, which must allow any skilled third party to implement the model. This is usually achieved by including formulas, flowcharts and example source code or pseudo code in the Recommendation. Building an independent implementation is often a very complex task, but the Recommendation should contain all the necessary information and the included conformance test data help to validate such independent implementations. Due to the exact specification, users of such models can trust the results of their measurements and can compare results across different measurement equipment from different vendors implementing one of these models and any implementation has an exactly known performance and scope. It should be noted that models described in such Recommendations may well be based on machine learning or artificial intelligence. The only thing which matters is the exact definition of the model and all its associated coefficients, etc.

Recommendations based on model descriptions, especially if they contain exactly one model description are the preferred solution for ITU-T Recommendations containing objective models for the assessment of perceptual quality factors.

Example of ITU-T Recommendations under the responsibility of SG12 falling into this category are (non-comprehensive list):

- ITU-T P.56 (method B)
- ITU-T P.79
- ITU-T P.502
- ITU-T P.563
- ITU-T P.700
- ITU-T P.862
- ITU-T P.863
- ITU-T P.1201.x
- ITU-T P.1202.x
- ITU-T P.1203.x
- ITU-T P.1204.x

6.1.2 Requirement specifications

NOTE – Please note that the term requirement specification in the context of this Supplement refers to a type of Recommendation and not to the document which is often used to describe work items to be developed by a Question.

Requirement specifications do not include model descriptions. Instead, they describe a scope and applications for a model. Instead of a model description, the Recommendation includes test vectors and tight bounds within which any implementation must reproduce these. Instead of providing test vectors as such, a method to create these test vectors can be provided as well. This method must be

described detailed enough to allow for the exact (within reasonable numerical accuracy) reproduction of the test vectors in different places and on different platforms.

Different implementations conforming to such a requirement specification are not exactly defined and results achieved with one implementation may differ significantly from results achieved by another implementation, even though both implementations may meet the requirements laid out in the Recommendation. Therefore, requirement specifications should demand that implementations are independently validated and become a normative part of a Recommendation after their validation.

Requirement specifications are a good solution if the scope of the Recommendation is very narrow and can be well covered with a set of test vectors. In any case, care should be taken, that enough test vectors are included and that the recommended limits for conformance testing are very tight.

Note that models which fulfil the requirements specified in such Recommendations may or may not be based on machine learning or artificial intelligence.

Example of ITU-T Recommendations under the responsibility of SG12 falling into this category are (non-comprehensive list):

- ITU-T O.41/P.53
- ITU-T P.56 (approximate equivalents of method B)
- ITU-T P.561
- ITU-T P.564
- ITU-T P.565
- ITU-T P.931

6.1.3 Recommendations specifying methods to develop measurement models

These specifications differ significantly from those in clause 6.1.2, in so far, as no conformance test vectors, or no detailed scope (of the resulting models) are included. The justification for such descriptions is to provide tools which can be adapted to yet unknown use cases. Instead of a model description, a detailed description of steps to be followed in order to create a final model is provided. There could also be the possibility that no conformance test vectors are provided, but instructions on how to generate new conformance test vectors for the final use case are provided.

The difference compared to requirement specifications is that the use case is not defined in the Recommendation. However, instructions are given on how to develop and test a model for a future use case and how test vectors can be created, once the use case is defined.

Due to the nature of these Recommendations, it is more difficult to guarantee that two implementations for the same use case provide the same results (within statistical significance) and thus a comparison of the measured results would be more valid, than is the case with model descriptions.

Also, one resulting model may be simple, efficient and accurate, while another model for the same use case, which was also developed according to the same method is totally over-trained and not generalizing well at all. This difference in accuracy forms a certain risk for non-expert users. Methods exist which can be applied to check for overfitting, but this demands independent validation to guarantee that these methods were applied properly, and the results of the checking were taken into account.

Without clear rules for independent validation of models developed following their content, Recommendations specifying methods to develop measurement models present a clear risk that control over the performance of the actual implementation can be taken out of the hands of the ITU-T, while such an implementation still bears the "ITU approved" mark. To leverage this risk, independent validation of the final, trained models must be a mandatory part of such Recommendations.

There are currently no Recommendations under the responsibility of SG12 falling into this category.

6.2 Special considerations regarding ML/AI based models

A typical ML/AI based model, which may be part of a Recommendation of any of the aforementioned categories, has several hundreds or even thousands of parameters which need to be optimized. Typically, during the training of the model, these parameters are iteratively varied and optimized based on comparison to the *learning* data set. After several iterations, the result is verified by comparison to a *validation* data set. Finally, after the training, the model performance is verified using a separate *test* data set. If the performance on the test data set proves to meet predefined requirements, the model can then be applied to real data and will not be adapted any further.

Depending on the structure of the model and the combination of test cases available in the training and validation set, such models may show very unexpected behaviour on data which were not included in these data sets.

At the core, classification trees (e.g., random forest) and artificial neural networks (ANN) are methods of classification. For easier understanding on how classification methods can be used for regression, i.e., to derive continuous values such as, MOS scores, the concept of a random forest is explained in a very simplified form as follows:

A random forest consists of many decision trees. Each tree is different and will classify the input data to a bin. Due to the difference between the trees, the resulting bins of each tree may also be different. Different trees in the forest thus come to slightly different results and a suitable aggregation of all these results (e.g., averaging, interpolation, …) leads to a pseudo continuous scale for the result.

In the case of voice quality measurements for example, the input could be certain parameters and/or statistics of parameters describing the distortions and the bins would be discrete MOS values.

During the learning phase of the random forest, the decisions of each tree are learned. The theory is that after the learning process, unknown input data are similar enough to the learning data, so that the aggregated decisions of all trees will be valid for these data too. The theory stands, depending on how well the learning data were selected against field data, and thus on how such data has been gathered. It is essential, that the entire scope of the use case is covered by the learning data. It also has to be taken into account, that an algorithm which does not try to mimic human perception sees the data differently than human beings. While human beings would typically classify distortion types by their perceptual impact or their technical cause, an algorithm will only "see" different input patterns. Distortions which human beings may see as equivalent, may however result in quite different input patterns for a measurement algorithm. Therefore, regardless of whether a perceptual or non-perceptual model is to be developed, it is the input patterns to the model which should be covered in their entirety by the training data. Of course, completeness will in most cases not be feasible and the coverage of the model input parameter space by training data will remain a compromise.

Due to their classification capabilities, it is also not uncommon, that for new input data, which were not seen in the training, AI or ML based models may react non-monotonic in an undesired way, and may show very strong, volatile variations of their results, even for very small changes of their input. There may also be excellent ML or AI based models, but to exclude the bad cases, the test data set must be much larger than for algorithmic models and the distortion ranges must be probed with a much finer granularity.

A consequence of this potential non-monotonicity is that two models trained on the same data and achieving very similar results on a validation data set may perform very different on other data. Long short-term memory (LSTM) and recursive networks may diverge even more.

It should also be noted, that after a test dataset has been used once, it is "burned", since from then on it must be seen as an a-priori known dataset. Datasets included in a Recommendation can only be used as learning or validation data but can never again be used as test datasets.

One type of AI or ML based technology uses a feature called continuous or self-learning. Models based on this concept will continuously try to improve their prediction based on feedback regarding

the correctness of previous predictions and thus, they will never reach a really stable state. Theoretically, they improve their accuracy and extend their applicability over time. However, there is no guarantee that the accuracy really improves and does not deteriorate, especially, when the feedback is in any way inaccurate or inconsistent. Also, two implementations of the same model will no longer produce the same scores for the same input after they had been applied in the field for a certain period. If such methods shall be applied to the measurement of quality factors, a reliable method needs to be found which can provide the required correct feedback, which is very difficult if that feedback should ultimately come from human beings. So far, it is unclear if such models will ever be developed and proposed for the measurement of quality factors. It may change in the future, but until further studies prove the opposite, continuous learning methods should be considered as non-verifiable, i.e., non-standardisable.

6.3 Guidelines for models

ITU-T has a long history of recommendations specifying measurement methods. Common to all these methods is the high quality of their specification and their careful validation. Independent reviews and validation tests make sure that any measurement method standardized by a Study Group has been rigorously tested and produces highly accurate results. Conformance test vectors ensure that users of conformant implementations can trust their results and will be able to compare results produced by different implementations. The following guidelines are widely applicable and used in SG12. They form the basis for the high acceptance of Recommendations under the responsibility of the Study Group:

1) Measurement results should be deterministic. This means, within the numerical accuracy (Note: numerical, i.e., repeatability, not absolute accuracy) of one implementation of the model, the same input must always produce the same output. For models measuring quality factors using the five-point MOS scale, a numerical accuracy of e.g., 0.0001 should be expected as a minimum. It should be noted that different models and applications may require different thresholds.

2) Generally, measurement results should be monotonic if perceptual degradations are monotonically increased (for degradations within the scope of the model).

3) Measurement results should be accurate. To prove this, a very large dataset with known correct measurement results is required ("ground truth"). This ground truth should ideally be the result of standardized subjective experiments.

4) The ground truth should be diverse, covering the entire range of distortions for which the method shall be standardized (i.e., the scope of the Recommendation) as well as many combinations of these distortions. The distortions and their combinations shall be adapted to real world applications. Testing unrealistic combinations is as useless as it is negligence to leave out realistic test conditions.

5) The test conditions included in the ground truth should not be decided by the model developers only, but also by including the expertise of potential users of the method.

6) The ground truth used for the validation of models should be large and diverse enough. How large exactly depends on the degree of freedom of the method and the diversity of input data for which the method was designed. Generally, depending on the ML/AI technology used, the number of samples required could be estimated based on the results of overfitting/underfitting tests (see guideline No. 8).

7) At least a significant part of the ground truth shall not be known to the developers at the time a method is developed. This part of the ground truth shall be used for an independent validation of the method before accepting it. Ideally, the unknown part is generated after freezing the method. The unknown part of the ground truth must span the entire scope of the model and ideally exceed this slightly in order to characterize the model performance at the borders of its scope.

8) Regardless of the model type, overfitting/underfitting needs to be tested. If ML/AI is used, then mathematical known techniques need to be applied.

9) The accuracy with which a method shall reproduce the results of the ground truth shall be defined by the Study Group. It is essential that the accuracy meets field test requirements.

10) A large set of conformance test vectors is required and for these test vectors, all implementations of the method should yield the same results within very strict bounds. The conformance test shall be a mandatory part of the Recommendation and shall be defined by the Study Group. The conformance test vectors should ideally span the entire scope of the model. If instead of test vectors, instructions are provided on how to generate such test vectors, these instructions shall be complete, and it should be possible to exactly reproduce the test vectors within reasonable numerical accuracy. Implementations of a model should reproduce the target values of the conformance test with very high accuracy, e.g., as defined in Annex A of [b-ITU-T P.863] (whereas it has to be noted, that practical implementations of ITU-T P.863 fulfill the strictest given bound in 100% of the cases) in different environments and on different platforms. Note that different models and applications may require different thresholds.

6.4 Guidelines for Recommendations

Based on the previous clauses, the following guidelines are established for new ITU Recommendations (in the context of SG12) targeting measurement methods.

A. Any model to be recommended should fulfil the guidelines (1to10) outlined in clause 6-3 and should have been independently validated accordingly.

B. Whenever possible, Recommendations should clearly define well specified measurement models, ideally a single unique model only.

C. For the time being, models based on continuous learning should be entirely avoided. Once, SG12 agrees that accurate enough feedback of ground truth may be produced and that training methods which are continuously applied are stable enough, this can be revised.

D. Methods to develop models are per se not specific enough to form a Recommendation, violate Annex D of the "Author's guide for drafting ITU-T Recommendations" and should thus be avoided. The *product* of such methods, i.e., a trained model, can however be well recommended (if other guidelines are fulfilled). The methods themselves are nevertheless valuable information and should be published as e.g., a Supplement or a Technical Report.

E. Recommendations containing requirement specifications can be accepted under the condition that their scope is narrow and can be well covered with conformance test vectors.

F. Recommendations with the same scope, the same technical requirements (e.g., required inputs), similar accuracy and no other clearly beneficial advantage compared to each other should be avoided. The same holds for Recommendations whose scope is a subset of an already existing Recommendation if no advantage over the existing Recommendation is provided. This shall be interpreted in the sense that redundant Recommendations should be avoided.

G. Before drafting a Recommendation, it should be verified that it will contain material which can be normative. Recommendations which do not describe procedures, best practices, algorithms or thresholds, or which do not define terms shall be avoided (see WTSA Res. 1 (1bis.5.1)). Findings directly related to an existing Recommendation can be published in a Supplement.

Bibliography

[b-ITU-T P.10] Recommendation ITU-T P.10 (2017), *Vocabulary for performance, quality of service and quality of experience*.

[b-ITU-T P.863] Recommendation ITU-T P.863 (2018), *Perceptual objective listening quality prediction*.

[b-ML] Wikipedia contributors (2019, November 4), Machine learning. In *Wikipedia, The Free Encyclopedia*. <https://en.wikipedia.org/w/index.php?title=Machine_learning&oldid=924580008>

SERIES OF ITU-T RECOMMENDATIONS

Series A	Organization of the work of ITU-T
Series D	Tariff and accounting principles and international telecommunication/ICT economic and policy issues
Series E	Overall network operation, telephone service, service operation and human factors
Series F	Non-telephone telecommunication services
Series G	Transmission systems and media, digital systems and networks
Series H	Audiovisual and multimedia systems
Series I	Integrated services digital network
Series J	Cable networks and transmission of television, sound programme and other multimedia signals
Series K	Protection against interference
Series L	Environment and ICTs, climate change, e-waste, energy efficiency; construction, installation and protection of cables and other elements of outside plant
Series M	Telecommunication management, including TMN and network maintenance
Series N	Maintenance: international sound programme and television transmission circuits
Series O	Specifications of measuring equipment
Series P	**Telephone transmission quality, telephone installations, local line networks**
Series Q	Switching and signalling, and associated measurements and tests
Series R	Telegraph transmission
Series S	Telegraph services terminal equipment
Series T	Terminals for telematic services
Series U	Telegraph switching
Series V	Data communication over the telephone network
Series X	Data networks, open system communications and security
Series Y	Global information infrastructure, Internet protocol aspects, next-generation networks, Internet of Things and smart cities
Series Z	Languages and general software aspects for telecommunication systems

Printed in Switzerland
Geneva, 2020