ETSI TR 103 559 V1.1.1 (2019-08)

Speech and multimedia Transmission Quality (STQ); Best practices for robust network QoS benchmark testing and scoring

Reference
DTR/STQ-00219m

Keywords
3G, benchmarking, data, GSM, LTE, network, QoE, QoS, scoring, service, speech, video, ViLTE, VoLTE

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

The present document can be downloaded from:
http://www.etsi.org/standards-search

The present document may be made available in electronic versions and/or in print. The content of any electronic and/or print versions of the present document shall not be modified without the prior written authorization of ETSI. In case of any existing or perceived difference in contents between such versions and/or in print, the prevailing version of an ETSI deliverable is the one made publicly available in PDF format at www.etsi.org/deliver.

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
https://portal.etsi.org/TB/ETSIDeliverableStatus.aspx

If you find errors in the present document, please send your comment to one of the following services:
https://portal.etsi.org/People/CommiteeSupportStaff.aspx

Contents

Intellectual Property Rights

Essential patents

IPRs essential or potentially essential to normative deliverables 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 (https://ipr.etsi.org/).

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.

Trademarks

The present document may include trademarks and/or tradenames which are asserted and/or registered by their owners. ETSI claims no ownership of these except for any which are indicated as being the property of ETSI, and conveys no right to use or reproduce any trademark and/or tradename. Mention of those trademarks in the present document does not constitute an endorsement by ETSI of products, services or organizations associated with those trademarks.

Foreword

This Technical Report (TR) has been produced by ETSI Technical Committee Speech and multimedia Transmission Quality (STQ).

Modal verbs terminology

In the present document "**should**", "**should not**", "**may**", "**need not**", "**will**", "**will not**", "**can**" and "**cannot**" are to be interpreted as described in clause 3.2 of the ETSI Drafting Rules (Verbal forms for the expression of provisions).

"**must**" and "**must not**" are **NOT** allowed in ETSI deliverables except when used in direct citation.

Introduction

Countrywide mobile network benchmarking and scoring campaigns published in the press enjoy great public interest and are of high importance for the operators of mobile networks. A first place score in press releases associated with such measurements is often used in the advertisements of the winning operator to boost their corporate identity. Though published results are often well documented, they are not always completely transparent about how the actual scoring has been achieved. Methods and underlying assumptions are mostly not described in detail.

The present document discusses the construction and methods of such a countrywide measurement campaign, with respect to the area and population to be covered, the collection and aggregation of the test results and the weighting of the various aspects tested. The applicability of the results of such a campaign, for inter country comparison purposes, is not covered in the present document.

Based on established methods and quality metrics, such as success ratio and setup times, the results of the data collected in the benchmarking are aggregated individually. The individual aggregated values are weighted and further aggregated for each application like telephony, video and data services. The application fields are then in turn weighted and aggregated over the different areas where the data is collected. Finally, calculation of an overall score or a joint score is performed.

The experienced quality of service varies over time so that the individual score of a particular throughput cannot be fixed once and for all. As well as the test metrics changing over time, so does the importance of the various services. The present document describes a typical set of tests that could be performed and a related evaluation criteria. In the annexes, actual real-world examples of weightings and score mapping parameters are given.

1 Scope

The present document describes the best practices for benchmarking of mobile networks. The goal of the benchmarking is to determine the best provider or operator for a designated area with respect of the services accessed with a mobile phone. The tests conducted are telephony, video streaming, data throughput and more interactive applications such as browsing, social media and messaging. This goal is achieved by executing benchmarking tests in designated test areas that represent or actually cover a major part of the users of mobile services. The results collected in the various areas are individually and collectively weighted and summarized into an overall score.

Due to the rapid development of the mobile technology and consumption habits of the users, the quality of experience of the users changes over time even when the objective to measure the quality of service does not change. The present document needs to keep up with those changes and does so by parameterizing the individual factors that contribute to the score.

2 References

2.1 Normative references

Normative references are not applicable in the present document.

2.2 Informative references

References are either specific (identified by date of publication and/or edition number or version number) or non-specific. For specific references, only the cited version applies. For non-specific references, the latest version of the referenced document (including any amendments) applies.

NOTE: While any hyperlinks included in this clause were valid at the time of publication, ETSI cannot guarantee their long term validity.

The following referenced documents are not necessary for the application of the present document but they assist the user with regard to a particular subject area.

[i.1] ETSI TS 102 250-2: "Speech and multimedia Transmission Quality (STQ); QoS aspects for popular services in mobile networks; Part 2: Definition of Quality of Service parameters and their computation".

[i.2] Void.

[i.3] ETSI TR 102 505: "Speech and multimedia Transmission Quality (STQ); Development of a Reference Web page".

[i.4] ETSI TR 101 578: "Speech and multimedia Transmission Quality (STQ); QoS aspects of TCP-based video services like YouTube™".

[i.5] ETSI TR 102 678: "Speech and multimedia Transmission Quality (STQ); QoS Parameter Measurements based on fixed Data Transfer Times".

[i.6] ETSI TR 103 138: "Speech and multimedia Transmission Quality (STQ); Speech samples and their use for QoS testing".

[i.7] Recommendation ITU-T E.840: "Statistical framework for end-to-end network-performance benchmark scoring and ranking".

[i.8] Recommendation ITU-T P.1401: "Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models".

[i.9] Recommendation ITU-T P.863: "Perceptual objective listening quality prediction".

[i.10] Recommendation ITU-T P.863.1: "Application guide for Recommendation ITU-T P.863".

3 Definition of terms, symbols and abbreviations

3.1 Terms

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

live web page: web pages considered as dynamic content, content changes over time and some content might be different caused by the hosting server or the access network

static web page: web pages considered as static content, content stays constant over time and access network

3.2 Symbols

Void.

3.3 Abbreviations

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

AMR	Adaptive Multi-Rate
API	Application Programming Interface
CDN	Content Delivery Network
CST	Call Setup Time
DL	DownLink
DNS	Domain Name System
EVS	Enhanced Voice Services
FB	FullBand
HD	High Definition
HTTP	HyperText Transfer Protocol
IP	Internet Protocol
ITU	International Telecommunication Union
ITU-T	International Telecommunication Union Telecommunication
KPI	Key Performance Indicator
MB	MegayByte
MOS	Mean Opinion Score
SMS	Short Messaging Service
TS	Technical Specification
UL	UpLink
VSSSR	Video Streaming Service Success Ratio
WB	WideBand

4 Governing Principles for Mobile Benchmarking

4.1 General

The accurate benchmarking and scoring of networks which cover large geographic areas requires careful consideration of a number of factors. These include the technology used, the extent of coverage offered, mobile device evolution, customer population distribution, network usage and tariff offerings. The following principles should be adhered to where possible to ensure that benchmarking scoring outcomes are always meaningful.

4.2 Fair Play

Benchmarking outcomes can be significantly influenced by specific targeting of test devices for superior performance. In such cases the results obtained no longer reflect the experience of a customer using that network. Steps should be taken to ensure that the measured results are truly representative of the real customer experience.

For example, if Operator A implements a special QoS construct specifically for the devices used to collect Benchmarking data, and Operator B does not, the results should not be compared for the purpose of drawing conclusions about the relative experience of customers on each network. The networks should not be compared for benchmarking purposes.

For example, if Vendor A implements a special functionality in their equipment/device software or firmware to recognize benchmark testing and boost performance, and Vendor B does not, the results may show one vendor to be superior to another for test cases no longer relevant to usual network usage. Vendor performance, from a customer perspective, can no longer be reliably compared.

4.3 Comparing networks with different coverage extents

Often networks are built with differing coverage objectives. Network rollout often varies between operators. This is often an important differentiator for customers making decisions about which network is best for them. Benchmarking should be performed in such a way that it highlights coverage differences in the results. From a scoring perspective, operators should never be penalized for providing coverage where other operators do not. In fact they should instead be rewarded in the scoring system. It should be the intention of any comprehensive mobile benchmark to include coverage comparison as a differentiating factor in the scoring.

For example, if Operator A offers significantly more geographic coverage than Operator B, Benchmarking data collection methodology and scoring should be such that this difference is always reflected in the scoring as a 'bonus' rather than a 'penalty' and the Benchmarking methodology should be such that this difference is measured. Failures occurring due to lack of coverage should always be included in scoring calculations and weighted appropriately to reflect the true customer experience.

4.4 Comparing networks with differing technology use

Network evolution and the adoption rate of new technologies often varies between operators. Benchmarking should be performed in such a way that it incorporates the use of the latest technology available. This is to reflect the network capability and customer experience available with the latest devices. Benchmark scoring should account for Operators who offer performance differentiation through early adoption of new technologies by way of a 'bonus' for such deployment.

For example, if operator A deploys 5G technology whilst operator B continues to deploy 4G technology, the benefits 5G technology offer to the customer experience should be captured in the Benchmarking data collection and scoring.

4.5 Test device selection

Mobile network benchmarking is performed mainly using drive testing. This relies heavily on the choice of test device(s). Care should be taken in the selection of such devices to ensure they do not favour one Operator's network over another in the results. The same devices may perform differently on two different networks depending on factors such as the antenna placement in the device for varying frequency bands, variations due to manufacturing tolerances, firmware version differences, modifications made to devices for metric data collection and device placement and mounting in the test vehicle.

4.6 Test server selection

Data tests are commonly performed to a test server or selected web page (or pages). The selection of such servers/sites can influence the benchmarking result. Test servers should be selected so they do not favour one network compared to another. Web pages should be selected such that they represent a cross section of pages commonly used by customers.

For example, if Operator A hosts the sever selected for 'ping' testing and the same server is also used to test Operator B, it is likely that performance levels for Operator B will be worse than those for Operator A due to the difference in latency to the selected server. This miss-represents the performance difference for this metric. Such situations should be avoided.

4.7 Test method transparency

Given the importance of the clear interpretation of benchmark results, all results should be accompanied by a declaration containing information about the following:

1) The scoring model/methodology used including all coefficients, targets and weightings.
2) The underlying KPI values as measured in the test.
3) The number of samples collected or number of tests performed for each KPI measured for each sub category.
4) The test methodology used including details of equipment setup, call sequences, test servers and web pages.
5) The areas/routes used for the data collection.
6) The device model and firmware version used for the data collection.
7) The tariff/data plan used for the data collection.

The intention of this is to provide the transparency required so that parties receiving the results are able to understand them fully. All factors required for this understanding should be provided.

4.8 Advice and best practice for web-page selection

Web page selection can impact on webpage load test results. To ensure a representative performance comparison can be made the following information and advice should be considered:

- For sufficient diversity and robustness of results, a minimum of 6 different pages is recommended to be considered for the scoring. It is good practice to measure more pages (e.g. 10), to retain enough diversity in case the dynamic behaviour requires to eliminate certain pages from the overall result.
- It is recommended to select pages according to their relevance to end customers. A popular ranking per country is given by Alexa Internet, Inc. (www.alexa.com). If possible, pages should be selected from Top 50 list, where an extension of that range is justifiable if not enough suitable pages exist within the Top 50.
- All pages should exceed a minimum size (e.g. 800 kB) to cover the minimum amount of data in case the download of a predefined data amount is used as success criteria. The page size needs to be observed on a daily basis throughout the measurements. In case of the severe size changes, a reaction may be needed.
- Internationally popular live pages and country dependent pages may be used in reasonable proportion (e.g. 10 live pages - 4 are common, 6 are country dependent).
- Ad blockers should not be used.
- A web-page selection that is hosted pre-dominantly by one CDN should be avoided.
- Websites of services that are predominantly accessed via a dedicated app on a smartphone should not be selected. For example, Facebook™, YouTube™ and similar websites/services are typically not accessed via a mobile browser and should therefore not be used as web-sites for HTTP Browsing tests in mobile benchmarking campaigns.
- No website should be selected that is a sub-page/site of another already selected website.
- No website should be selected where the content is legally suspicious or contains harming, racism or sexist content.

5 General Description

In the present document the benchmarking and scoring of networks over a large geographical area, e.g. entire countries in various modes and for diverse services provided by mobile networks is described. A comprehensive manner to compare the tested networks is to calculate an overall score per network based on the individual measurement results collected during a test campaign. The individual measurement results are aggregated using a weighted accumulation into an overall network score. This overall score finally allows the scoring of the tested networks. To arrive there, the weighted aggregation is done over several layers.

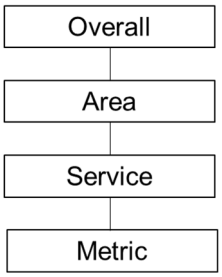

Figure 1: Aggregation layers

Weights are used for the aggregation of the different metrics, mobile services and areas to obtain the final score.

The accumulation of the measurements is done over several levels. The first or lowest layer consists of the measurement metrics for the services delivered over the mobile network. The services or applications considered are telephony, video, data transfer and services including browsing, social media and messaging. The metrics collected for one mobile service and a certain area are aggregated into an individual score for each metric; the scores of the metrics are then aggregated into an overall score of the mobile service.

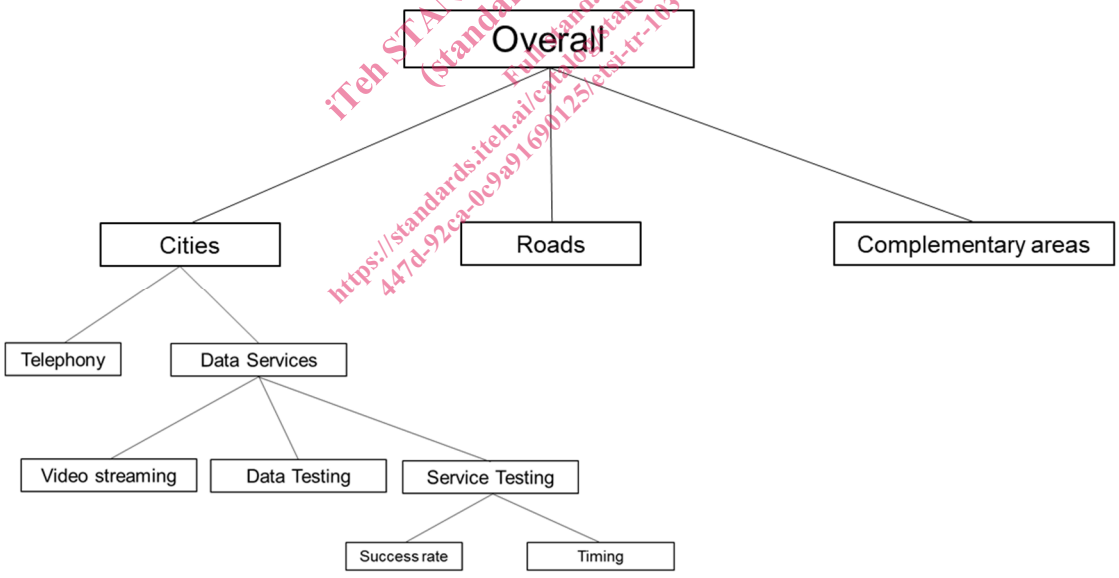

Figure 2: Aggregation over services and layers for a mobile network

In this aggregation, the metrics have a score weight according to the weight they were given for that particular mobile service. The scores for the individual mobile services are then in turn aggregated into a score for telephony and data services, and then together for the area they were collected in.

Finally the various areas are weighted and accumulated over the various areas covered in the measurement. The different areas can have further geographical subdivisions. The weighted aggregation of the areas results in an overall score that characterizes the network.