Final draft ETSI ES 202 765-2 V1.1.3 (2010-03)

ETSI Standard

Speech and multimedia Transmission Quality (STQ); QoS and network performance metrics and measurement methods; Part 2: Transmission Quality Indicator combining Voice Quality Metrics

Reference

RES/STQ-00148-2

Keywords

performance, QoS, user, voice

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

Contents

Intellectual Property Rights

IPRs essential or potentially essential to the present document 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 (http://webapp.etsi.org/IPR/home.asp).

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.

Foreword

This ETSI Standard (ES) has been produced by ETSI Technical Committee Speech and multimedia Transmission Quality (STQ), and is now submitted for the ETSI standards Membership Approval Procedure.

1 Scope

The present document aims at identifying and defining indicators and methodologies for a use in a context of end-user quality characterization and supervision of voice telephony services.

In this context the measurements and metric determinations are perform by analysing signals accessible on user-end services and not on the network. In order to mirror the reality in terms of access to the services at the user-end measurements and analysis are perform on electrical signal that exclude the electro-acoustic part of the end equipment but the probe adaptation to electric interface of the end user equipment much take into account the electro-acoustic characteristics of this terminal.

All the indicators presented and defined in the present document are objective indicators obtained by instrumental measurement methods.

2 References

References are either specific (identified by date of publication and/or edition number or version number) or non-specific.

- For a specific reference, subsequent revisions do not apply.
- Non-specific reference may be made only to a complete document or a part thereof and only in the following cases:
  - if it is accepted that it will be possible to use all future changes of the referenced document for the purposes of the referring document;
  - for informative references.

Referenced documents which are not found to be publicly available in the expected location might be found at http://docbox.etsi.org/Reference.

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

2.1 Normative references

The following referenced documents are indispensable for the application of the present document. For dated references, only the edition cited applies. For non-specific references, the latest edition of the referenced document (including any amendments) applies.

Not applicable

2.2 Informative references

The following referenced documents are not essential to the use of the present document but they assist the user with regard to a particular subject area. For non-specific references, the latest version of the referenced document (including any amendments) applies.

[i.1] ITU-T Recommendation P.800: "Methods for subjective determination of transmission quality".

[i.2] ITU-T Recommendation P.862: "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs".

[i.3] ITU-T Recommendation P.862.1: "Mapping function for transforming P.862 raw result scores to MOS-LQO".

[i.4] ITU-T Recommendation P.862.2: "Wideband extension to Recommendation P.862 for the assessment of wideband telephone networks and speech codecs".

[i.5] ITU-T Recommendation P.862.3: "Application guide for objective quality measurement based on Recommendations P.862, P.862.1 and P.862.2".

[i.6] ITU-T Recommendation P.800.1: "Mean Opinion Score (MOS) terminology".

[i.7] ITU-T Recommendation E.800: "Terms and definitions related to quality of service and network performance including dependability".

[i.8] ITU-T Recommendation E.845: "Connection accessibility objective for the international telephone service".

[i.9] ETSI EG 201 769: "Speech Processing, Transmission and Quality Aspects (STQ); QoS parameter definitions and measurements; Parameters for voice telephony service required under the ONP Voice Telephony Directive 98/10/EC".

[i.10] ITU-T Recommendation P.56: "Objective measurement of active speech level".

[i.11] ITU-T Recommendation O.41: "Psophometer for use on telephone-type circuits".

[i.12] ITU-T Recommendation G.131: "Talker echo and its control".

[i.13] ITU-T Recommendation G.168: "Digital network echo cancellers".

[i.14] ITU-T Recommendation G.114: "One-way transmission time".

[i.15] ITU-T Recommendation P.505: "One-view visualization of speech quality measurement results".

[i.16] ETSI EG 201 377 (all parts): "Speech Processing, Transmission and Quality Aspects (STQ); Specification and measurement of speech transmission quality".

[i.17] ITU-T Recommendation H.323: "Packet-based multimedia communications systems".

[i.18] ITU-T Recommendation H.225.0: "Call signalling protocols and media stream packetization for packet-based multimedia communication systems".

[i.19] ITU-Recommendation P.50: "Artificial voices".

[i.20] ITU-Recommendation P.501: "Test signals for use in telephonometry".

NOTE: This Recommendation includes an electronic attachment containing test signals for telephonometry applications.

[i.21] ETSI TR 102 506: "Speech Processing, Transmission and Quality Aspects (STQ); Estimating Speech Quality per Call".

3 Abbreviations

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

ADSL	Asymmetrical Digital Subscriber Line
ATA	Analog Telephone Adapter
IP	Internet Protocol
ISDN	Integrated Services Digital Network
ITU-T	International Telecommunication Union - Telecommunication standardization sector
GPS	Global Positioning System
GSM	Global System for Mobile communications
HATS	Head And Torso Simulator
MGCP	Media Gateway Control Protocol
MOS	Mean Opinion Score
MOS-LQOM	Mean Opinion Store-Listening Quality Objective Mixed bandwidths
PDD	Post Dialling Delay

PESQ	Perceptual Evaluation of Speech Quality
PSTN	Public Switched Telephone Network
RTP	Real Time Protocol
SIP	Session Initiation Protocol
UMTS	Universal Mobile Telecommunications Service
VoIP	Voice over Internet Protocol

4 Introduction

The assessment of transmission quality based on voice quality metrics is already addressed in several standards at ETSI (e.g. EG 201 377 [i.16] series) and elsewhere (mostly ITU-T recommendations from the P and G series). These different documents are addressing the measurement methodologies in terms of metrics, threshold, data acquisition or modelling of subjective opinion.

The objective of the present document is to complement this material with practical requirements of use in the context of service verification and benchmark on a large and representative scale from the point of view of the end-users or of the regulatory authorities. This has been made necessary by the current or recent evolutions of the telecommunication sector:

- the competitive environment, in particular in voice services, where public protocols with high quality services have been replaced by a multitude of service providers with less guarantees, and where clients can very easily change their service providers;
- the development of time varying quality in telecommunications, first in mobile offers (due to mobility and irregular network coverage), but now also for fix services (mostly VoIP);
- the cohabitation, interaction and competition between services based on different technologies.

Voice transmission quality is now recognized as a differentiating factor, but it remains very difficult to quantify.

To achieve the goal mentioned beforehand, there are several existing possibilities, not fully satisfying:

- Customer surveys. This is by far the cheapest way to assess the perception of end users. But the bias introduced by the other factors like price, as well as the fact that voice quality itself is rarely questioned as itself or in a satisfactory way (one never knows before a survey what are the problems encountered by end users), makes this source not really reliable.
- Pseudo-subjective tests, with a few human testers assessing the quality of real links in several situations. This method has the major drawback of its lack of reproducibility, and is often applied without using the standard metrics and quality scales that can be found in standards like ITU-T Recommendation P.800 [i.1]. It is also very long to run and not really cheap in the current competitive context where so many offers have to be assessed. And it is not easily applicable in a context of quality changing over time.
- Objective tests. This is the most reliable way, although it is also based on sampling and can cost a lot of money in the case of a large deployment of probes or robots.

The present document assumes that this last family of methodology answers the needs of a reliable comparison of telephony offers and is applied without combination with other methods.

What definitely matters is the point of view of the end-users. What they perceive is not only the result of the transmission of a signal across a network; the processing of this signal at the sending and at the receiving sides has also a big importance. Therefore, it seems obvious not to use passive network monitoring systems to assess end-to-end voice quality, but rather active systems simulating the behaviour of the end users, including the terminal. A big advantage of such an approach is that it is highly technical and protocol agnostic, and therefore compliant with the expectations of users, which are not judging voice quality of PSTN, GSM or VoIP services following different criteria.

Last important aspect that is addressed in the present document is the practical organization of measurement campaigns in order to get a realistic and reliable vision of the services as perceived by the end-users. In particular, the questions of the periodicity of measurement and of the geographical coverage (i.e. more generally the sampling approach).

In order to mirror the reality in terms of access to the services, a reliable measurement or supervision system should provide the possibility to collect information from probes or robots adapted to the most common interfaces available. This includes:

- analogue access (for the simulation of PSTN or of analogue phones behind an ATA box or an ADSL modem);
- ISDN access;
- handset (for any wireline terminal);
- electrical input and output (for PC soundcards of for any wireless terminal);
- GSM;
- UMTS;
- ethernet with IP phone termination (SIP, ITU-T Recommendation H.323 [i.17], MGCP, etc.).

Any combination of end-to-end connection between the types of access mentioned here have to be considered when a measurement campaign is scheduled. Nevertheless, of course, there are practical limitations:

- the number of measurements for a given type of access should be in proportion with its level of use in the real life;
- the number of probes and of measurement results available will be adapted to the real needs as well as to the capacity (mostly in terms of cost and of processing capability) of the entity running these measurements.

Figure 4.1 shows these different configurations and interfaces.

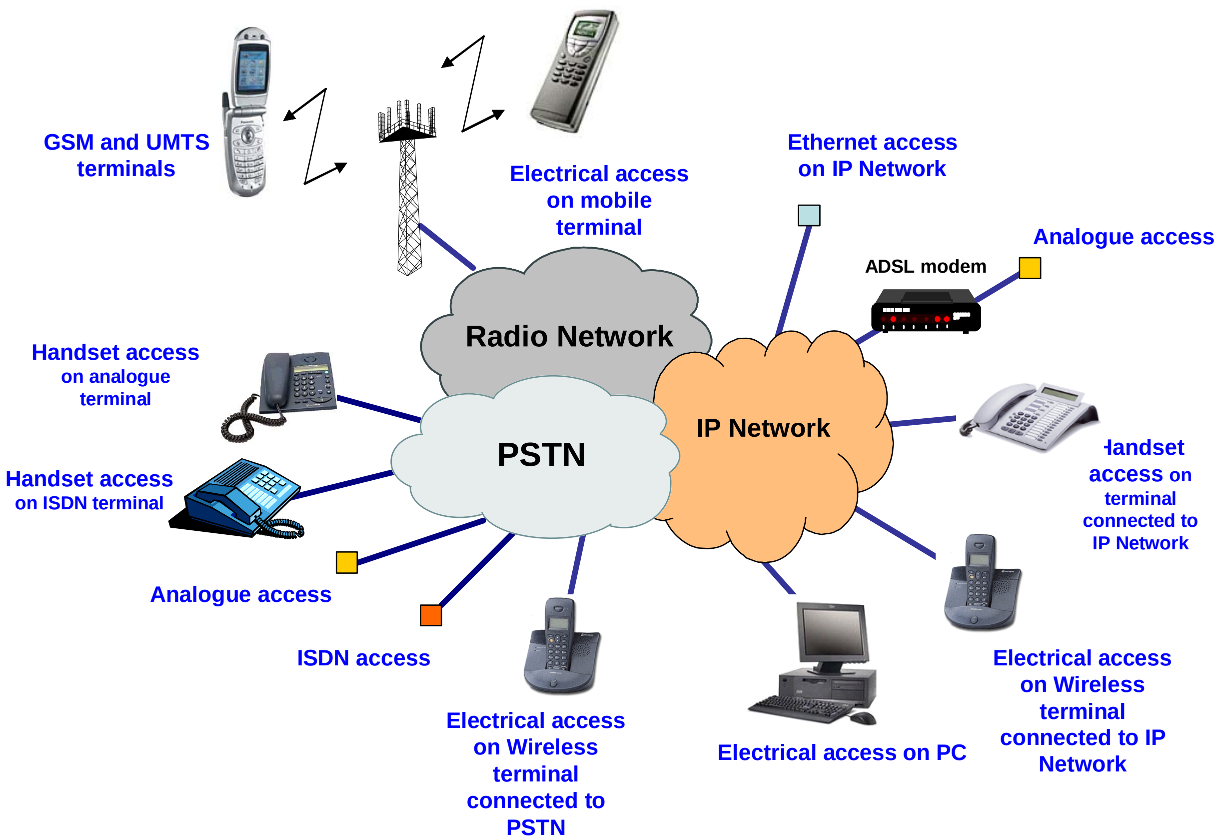


Figure 4.1: Possible configurations and interfaces in context of user characterization

5 Measurement type

To perform quality services assessments, there are two different methods: intrusive and non intrusive measurements.

The non intrusive measurements are not really adapted to end user surveys because it requires to install probes at the user's terminals.

The intrusive measurements are more adapted to end user surveys because probe connection with end user terminals is easier. Compared to non intrusive measurements, the intrusive methods have an advantage: the opportunity for voice quality assessment to use models with references such as ITU-T Recommendation P.862 [i.2] (see also ITU-T recommendations P.862.1 [i.3], P.862.2 [i.4] and P.862.3 [i.5] concerning mapping functions and application guide) which give results close to subjective perception of the speech quality.

In this context, the intrusive measurements using models working with references for speech quality assessment will be perform for end user survey.

6 Voice quality scale

It is important to consider that nowadays telephony has entered an era where traditional narrowband services will cohabit with new services offering wideband audio capacities. For end-users, these are not separated kinds of services. Therefore, the assessment of transmission quality of voice should now be based on common metrics and objective quality levels and scales, in replacement of the existing narrow-band only ones. In this context, it is appropriate to use the MOS-LQOM scale to characterize voice quality of narrow-band services and wideband services. See ITU-T Recommendation P.800.1 [i.6] for more information on MOS terminology.

7 List of indicators

The indicators for the context of end-user quality survey of voice services are given in the following sub-clauses.

It should be noted that the indicators defined below do not completely cover the conversational quality. Additional indicators may be:

- double talk performance parameters;
- switching characteristics parameters.

7.1 Post Dialling Delay

Definition	Post Dialling Delay (PDD) evaluates service availability to set up calls in an acceptable delay. It is linked to the service architecture complexity, and to the performance of the constituting network elements. Post Dialling Delay is the time interval between the end of dialling by the caller and the reception back by him of the appropriate ringing tone or recorded announcement. Metric determines on one of the two access of the communication.
Assessment method	Indicator determines sequentially from the two access of call configuration. This indicator characterizes only the caller part of the configuration.
Unit	Millisecond with an integer value.
Standardization reference	
Significant	Mandatory.
Comment	This indicator has to be separated between call types (IP to IP, IP to PSTN, IP to mobile, etc.) for a detailed analysis. The objective set up in for universal telephony service has been set up to 2 900 ms in the French regulator recommendation.

7.2 Media establishment delay

Definition	Time determines on one of the two access of the communication, between off hock of the called and the beginning of voice signal receive.
Assessment method	Indicator determines sequentially from the two access of call configuration. On an IP access this indicator may be assessed by using a non-intrusive probe, such as a protocol analyser. Media establishment delay may be evaluated through the analysis of media flows and signalling. For ITU-T Recommendation H.323 protocol [i.17] the flow establishment delay corresponds to the time elapsed between the emission of the ITU-T Recommendation H.225.0 [i.18] "CONNECT" message and the arrival of the first IP packet including speech signal.
Unit	Millisecond with an integer value.
Standardization reference	
Significant	Optional.
Comment	This indicator has to be separated between caller and called site for a detailed analysis.

7.3 Unsuccessful call ratio

Definition	Ratio of unsuccessful calls to the total number of call attempts in a specified time period. An unsuccessful call is a call attempt to a valid number, properly dialled following dial tone, where neither called party busy tone, nor ringing tone, nor answer signal, is recognized on the access line of the calling user within 30 seconds from the instant when the address information required for setting up a call is received by the network.
Assessment method	Indicator determines sequentially from the two access of call configuration.
Unit	% with the resolution of 1 digit after the decimal point.
Standardization reference	ITU-T Recommendation E.800 [i.7], ITU-T Recommendation E.845 [i.8], EG 201 769 [i.9].
Significant	Mandatory.
Comment	The limit of 30 seconds is the default set-up of a timer in SS7 protocol.

7.4 Premature release probability

Definition	This indicator characterizes the ability to release a service. It is based on the measurement of the number of released communications in comparison with the number of established communications. Released communications are defined as communications released before voluntary action from one of the ends of the transmission.
Assessment method	
Unit	% with the resolution of 1 digit after the decimal point.
Standardization reference	ITU-T Recommendation E.800 [i.7].
Significant	Optional.

7.5 Level of active speech signal at reception

Definition	Level of speech signal received after transmission. The level of the signal heard by the user has an impact on the quality he will perceive. A too low signal will be hardly audible and masked by the noise, while a too high level will be painful. Therefore, a measurement of the speech signal level is necessary to ensure a good listening comfort.
Assessment method	The received decoded signal used for instance for ITU-T Recommendation P.862 [i.2] can be used also to assess this parameter. A typical method for the measurement of this parameter, based on a sample by sample approach and a moving threshold between noise and speech, is given in ITU-T Recommendation P.56 [i.10].
Unit	dBm with the resolution of 1 digit after the decimal point.
Standardization reference	ITU-T Recommendation P.56 [i.10].
Significant	Optional.
Comment	Each sample of signal has a level, generally express in mV. The mean speech level is the transformation on as appropriate logarithmic scale of the mean signal voltage. The samples taken into account for this measurement are the ones seen as speech (the others are taken into account for noise measurements). It is recommended to fall within classical speech levels values, i.e. between -25 dBm and -10 dBm.

7.6 Noise level at reception

Definition	Level of noise determines at reception in non-speech segment of speech sample. The noise present besides the speech signal can have characteristics that can become a disagreement, for instance if they have a varying spectrum (crowd, noise, for instance). But the more important source of annoying due to noise is simply its level.
Assessment method	The received decoded signal used for instance for ITU-T Recommendation P.862 [i.2] can be used also to assess this parameter. The measurement of these parameters is normally performed as for speech signal level (see clause 7.5), but on the samples identified as non-speech.
Unit	dBmOp with the resolution of 1 digit after the decimal point.
Standardization reference	ITU-T Recommendation O.41 [i.11].
Significant	Optional.
Comment	Each sample of signal has a level, generally express in mV. The mean noise level is the transformation on as appropriate logarithmic scale of the mean signal voltage of the noise samples. To get a more accurate noise level measure, a frequency transform needs to be done in order to apply a *psophometric* weighting (see ITU-T Recommendation O.41 [i.11]). It is recommended not to have noises louder than -50 dBmOp.

7.7 Noise to signal ratio at reception

Definition	Difference between the active vocal level and the level of noise at the reception. The noise present besides the speech signal can have characteristics that can become a disagreement, for instance if they have a varying spectrum (crowd, noise, for instance). But the more important source of annoying due to noise is simply its level, and particularly the relative level compared to speech.
Assessment method	Combination of speech signal level (see clause 7.5) and noise level (see clause 7.6) can replace noise to signal ration indicator. The received decoded signal used for instance for ITU-T Recommendation P.862 [i.2] can be used also to assess this parameter. Speech signal level/noise level = SNR. Indicator determines in the two directions of transmission.
Unit	dB with the resolution of 1 digit after the decimal point.
Standardization reference	ITU-T Recommendation P.56 [i.10].
Significant	Optional.
Comment	It is recommended not to have SNR lower than 30 dB.

7.8 Speech signal attenuation (or gain) after transmission

Definition	Variant metric of the speech signal level (if one knows the sending speech level, they are even redundant). Speech signal attenuation after transmission is the difference between the active vocal level at receiving and sending access.
Assessment method	The received decoded signal used for instance for ITU-T Recommendation P.862 [i.2] can be used also to assess this parameter. Once the speech signal level has been computed (see clause 7.5), it is compared with the level of the sent signal. The attenuation is the difference between these two levels. There are other methods to compute this parameter, based for instance on intrusive measurement made with sine waves and a specific weighting function. Indicator determines in the two directions of transmission.
Unit	dB with the resolution of 1 digit after the decimal point.
Standardization reference	
Significant	Optional.
Comment	It is recommended to comply with PSTN attenuation rules, i.e. an attenuation between 6 dB and 10 dB.

7.9 Talker echo delay

Definition	In telecommunications, the term echo describes delayed and unwanted feedback of the send signal into the receive path. The so-called echo source is the reflection point between send and receive directions, which could be one of the following causes: • 4-wire/2-wire Hybrid Circuits (multiple reflections possible); • coupling in handset cords; • structure borne coupling in handsets; • acoustical coupling between earpiece and microphone. This phenomenon is characterized by two parameters: its attenuation and its delay. See annex C for a more detailed discussion of talker echo and the related listener echo. With the increased delays present in today's IP networks, echo has the potential to be much more perceivable and annoying than in classical PSTN. In order to achieve a similar user perception with higher delays the attenuation of the talker echo should be increased, i.e. active echo cancellation is necessary. In practice it can be observed that in some cases, either the cancelling does not occur, or it is not fully performing. Echo is characterized by two parameters: its attenuation and its delay. The less attenuation and/or the more delay, the more the echo will become annoying. Echo delay is the time it takes for the speech signal to go from the mouth of a subscriber back to the ear of the same subscriber, with one or more reflections occurring along the transmission path.
Assessment method	Indicator determines sequentially from the two access of call configuration.
Unit	Milliseconds with an integer value.
Standardization reference	ITU-T Recommendation G.131 [i.12].
Significant	Optional.
Comment	For fully digital networks the talker echo delay can be assumed to be equivalent to twice the mean one-way delay.

7.10 Talker echo attenuation

Definition	In telecommunications, the term echo describes delayed and unwanted feedback of the sent signal into the receive path. The so-called echo source is the reflection point between send and receive directions, which could be one of the following causes: • 4-wire/2-wire Hybrid Circuits (multiple reflections possible); • coupling in handset cords; • structure borne coupling in handsets; • acoustical coupling between earpiece and microphone. This phenomenon is characterized by two parameters: its attenuation and its delay. See annex C for a more detailed discussion of talker echo and the related listener echo. With the increased delays present in today's IP networks, echo has the potential to be much more perceivable and annoying than in classical PSTN. In order to achieve a similar user perception with higher delays the attenuation of the talker echo should be increased, i.e. active echo cancellation is necessary. In practice it can be observed that in some cases, either the cancelling does not occur, or it is not fully performing. Echo is characterized by two parameters: its attenuation and its delay. The less attenuation and/or the more delay, the more the echo will become annoying. Echo attenuation is the difference of level between the sending level and the (delayed) receiving level both measured at the same subscriber while he is talking.
Assessment method	Indicator determines sequentially from the two access of call configuration.
Unit	dB with a resolution of one decimal.
Standardization reference	ITU-T Recommendation G.131 [i.12], ITU-T Recommendation G.168 [i.13].
Significant	Optional.
Comment	If active echo cancellation is used, ITU-T Recommendation G.168 [i.13] should be applied; VoIP should provide echo attenuation of 55 dB. If the delay under all circumstances is known to not exceed 50 ms lower values for the talker echo attenuation (down to 35 dB) may be acceptable (see annex C). Echo annoyance depends on two metrics: the attenuation and the delay. For a similar attenuation level greater the delay is more important the annoyance will be for the user(s). Echo-Annoyance factor is defined as K. $K = EA - 40 \times \text{Log}\ [(1 + \text{delay}/10) / (1 + \text{delay}/150)] + 6 \times \exp(-0{,}3 \times \text{delay}^2)$. NOTE: It is recognized that, besides insufficient echo talker attenuation, short term echo artefacts may occur which are not covered by this measurement.

7.11 Listening speech quality

Definition	Represents the intrinsic quality of speech signal after transmission. This indicator takes into account the degradations generated on the signal by the transmission links.
Assessment method	Voice quality is evaluated by using the ITU-T Recommendation P.862 [i.2] with the mapping functions according to ITU-T Recommendation P.862.1 [i.3] and ITU-T Recommendation P.862.2 [i.4]. MOS-LQO or Mean Opinion Score - Listening Quality Objective (calculated using the Perceptual Evaluation of Speech Quality, or ITU-T Recommendation P.862 [i.2] in conjunction with ITU-T Recommendation P.862.1 [i.3]) provides an objective view on the quality of the voice signal as it may be perceived by the customer. The MOS-LQO score is obtained by comparing speech samples: • the original signal sent by the far end of the connection; • the degraded signal received at the local end, where the measurement is applied. Since during a call, speech quality can vary, several MOS-LQO scores should be determined in series during the same call. So for a given transmission way, listening speech quality performance during the call is defined by the mean value of MOS-LQOM measurements (in the same direction). The voice quality indicator should be determined in both transmission directions. It is important that speech quality analyses are performed for the same call according to the direction of transmission. It is also important that speech quality analyses are performed from the beginning to the end of the call, by alternating the direction of transmission for the MOS-LQO score determination. If the first evaluation is made in the direction A to B, the second evaluation should be realized in the direction B to A, the third in the direction A to B, the fourth in the direction B to A and so on. Hence the number of analyses will be equal for each transmission direction. The use of the ITU-T Recommendation P.862 [i.2] has some limitations. ITU-T Recommendation P.862.3 [i.5] provides some important remarks that should be taken into account in the objective quality evaluation of speech conforming to ITU-T Recommendations P.862 [i.2], P.862.1 [i.3] and P.862.2 [i.4]. Users of ITU-T Recommendation P.862 [i.2] should understand and follow the guidance given in this Recommendation. The listening speech quality is calculated as the mean value of the MOS-LQO as in TR 102 506 [i.21] separately for each transmission direction.
Unit	Note between 1 (= very bad) and 5 (= excellent) determines on MOS-LQOM scale with a resolution of two digits after the decimal point.
Standardization reference	ITU-T Recommendation P.800 [i.1], ITU-T Recommendation P.800.1 [i.6], ITU-T Recommendation P.862 [i.2], ITU-T Recommendation P.862.1 [i.3], ITU-T Recommendation P.862.2 [i.4], ITU-T Recommendation P.862.3 [i.5]. TR 102 506 [i.21]
Significant	Mandatory.
Comment	The value of this indicator depends on the used codec, but also on impairments like IP packet loss or low signal to noise ratio. To make easier result comparisons, it is recommended to use the same speech samples in all test configurations, or at least select different speech samples based on a thorough selection and validation process. ITU-T Recommendation P.862 [i.2] is measuring listening quality and does not take into account impairments affecting conversational quality, like the end-to-end delay or echo. It is also a one-way indicator, therefore it has to be measured separately in both transmission directions, with no average between them afterwards. This indicator may be separated between call types (IP to IP, IP to PSTN, IP to Mobile, etc.) for a detailed analysis. For the time being , ITU-T Recommendation P.862 [i.2] is the only relevant and available objective method for speech quality evaluation in listening condition. A new model (P.OLQA) is in development in ITU.

7.12 Listening speech quality stability

Definition	It is well known that for IP networks, delay within the network can vary significantly, due to congestion or indeed due to route changes during a session. Delay variations in the network can be compensated to some extent through good jitter buffer design in the receiver. Furthermore, packet loss in the network can occur due either to severe congestion (buffer overflows) or routing problems, or in a receiver terminal due to jitter buffer overflow. These factors will impact on the quality of the service. Concerning voice over IP, a single measurement of speech quality once at the very beginning of a call is not enough. They should be analysed all along the duration of the call, typically several minutes. This metric represents the stability of the voice quality during a communication of several minutes long. This indicator takes into account the degradations generated on the signal by the transmission links.
Assessment method	Several measurements of MOS-LQOM score performed with ITU-T Recommendation P.862 [i.2] and the adapted mapping function (see clause 7.11) are performed in series within the same call. Typically, a measurement each 20 s is enough. The results are reported in terms of statistics. The assessment of Listening Speech Quality Stability is performed in 5 steps. The generic formulation is presented in annex A. For stability indicator about Listening Speech Quality, THRESHOLD1 = 0,1 and the linear weighting function applies in order to express Stability (ST-MOS) on a 0 to 100 scale. By definition Stability equals 100 when no instability occurs and Stability ST-MOS equals 0 when instability is equal or more than 0,4. ST-MOS is calculated as: • ST-MOS = 100 - (250 × INS_MOS); and • ST-MOS = 0 if [100-(250 × INS_MOS)] < 0. This indicator is determined in the two directions of transmission.
Unit	Statistics on MOS-LQO score variation are plotted on a 0 to 100 scale.
Standardization reference	
Significant	Mandatory.
Comment	The determination of this indicator is meaningful only if MOS-LQO scores are determined from the beginning to the end of the call by making successive analyses for alternating transmission directions.

7.13 End to end delay

Definition	Represent the global delay from one access to the other one. This indicator takes into account the transmission delay on networks but also processing delay in sending and receiving terminals. End to end delay is one of the components of perceived voice quality, may be the major one in VoIP. A great delay impacts negatively on QoS perceived by customer.
Assessment method	The end to end delay is the delay from mouth to ear, which means the transmission delay over the whole transmission path. For the purpose of the present document, end to end delay does not take into account the transducers delay (loudspeaker and microphone) while measurements are done at the electrical interfaces of the end terminals. To measure end to end delay it is needed to ensure a synchronization of both transmission ends of the measurement device. This synchronization may be done by GPS clocks when the two ends are distant. When ends are co-located synchronization may be done directly by the analyser. To assess the metric, the clock accuracy of the analyzer (or two analyzer parts) should be better than 10 ppm. Since during the call, end to end delay can vary, several measurements of delay are performed in series during the same call. So for a given transmission way, end to end delay performance during the call is defined by the mean value of delay measurements (in the same direction). The end to end delay is determined in both directions of transmission. It is important that delay measurements are performed during the same call according to the direction of transmission. It is also important that delay measurements are performed from the beginning to the end of the communication, by alternating the direction of transmission. If the first evaluation of delay is performed in the direction A to B, the second evaluation should be realized in the direction B to A, the third in the direction A to B, the fourth in the direction B to A and so on. Hence the number of analyses will be equal for each transmission direction. The end to end delay is calculated as the mean value of measurements in each transmission direction.
Unit	Millisecond with an integer value.
Standardization reference	ITU-T Recommendation G.114 (session 1) [i.14].
Significant	Mandatory.
Comment	The metric measurement needs time synchronization between the two parts of the analyzer connected to the both access of the communication path. The recommendations (ITU-T Recommendation G.114 [i.14] in particular) indicate not going beyond 150 ms in one-way.

7.14 End to end delay variation

Definition	End-to-end delay variation can occur, principally caused by changes within the network but also by sender and receiver terminals (e.g. Jitter buffer). This will have an impact on the quality of the service. Concerning voice over IP, the single measurement of end to end delay once at the very beginning of a call is not enough. It should be analysed all along the duration of the call, typically several minutes. This metric defines the stability of end to end delay during a communication of several minutes.
Assessment method	Several measurements of delay (according to clause 7.13) are performed in series, or the delay is measured at every short period within the same call. The results are reported in terms of statistics. End to end delay variation is defined from an indicator of the stability of delay variation, over a single communication. As for listening speech quality, the calculation of this indicator is performed in 5 steps. The generic formulation is presented in annex A. For stability indicator about end to end delay, THRESHOLD1 = 5 and the linear weighting function applies in order to express Stability (ST-Delay) on a 0 to 100 scale. By definition Stability equals 100 when no instability occurs and Stability ST-Delay equals 0 as soon as instability reaches or is greater than 10. ST-Delay is calculated as the following linear weighting: • ST-Delay=100-(10 × INS_Delay), and • ST-Delay= 0 if [100-(10 × INS_Delay)] < 0. This indicator is determined in the two ways of transmission.
Unit	Statistics on delay variation are plotted on a 0 to 100 scale.
Standardization reference	
Significant	Mandatory.
Comment	The determination of this indicator is meaningful only if end to end delays are determined from the beginning to the end of the call by making successive analyses for alternating transmission directions.

7.15 Frequency responses at the reception

Definition	Narrow-band telephony should transmit signals between 300 Hz and 3 400 Hz. Wide-band telephony should transmit signals between 50 Hz and 7 000 Hz. The objective of this measurement is to see what bandwidth is used, and also to see whether a partial and unwanted bandwidth limitation is present. Frequency response is the gain (or attenuation) of the level of speech signal after transmission according to frequency.
Assessment method	The received decoded signal as defined in ITU-T Recommendation P.862 [i.2] can be used also to assess this parameter. In general, a simple spectral transformation of the signal is enough. The attenuation or level of the signal is then measured for a selected number of frequencies (or frequency bands). For the purpose of the present document the measurement of the receiving frequency response needs to take into account the characteristics of the terminals transducers (loudspeaker and microphone). The analyses are done at the electrical interfaces of the terminals, it is necessary to ensure a calibration of the frequency response of the transducers. This calibration offers the possibility to adapt the electrical measurement values and to assess the actual frequency response at the receiving end (see calibration method in annex B).
Unit	Boolean (pass or fail) compared to a mask or value representing size of the frequency band used.
Standardization reference	
Significant	Optional.

8 Measurement frequency

Measurements need to take into account the possible variations of quality versus time. In this context, it is important to adapt the measurement frequency with the possible variations of quality. So it does not seem reasonable to go beyond 10 to 15 minutes between two consecutive calls on the same connection. A measurement frequency of 4 or 5 analysis by hour is adapted to end-user quality survey of voice services. This recommended measurement frequency applies for all indicators defined in clause 7 of the present document.

9 Duration of test calls

Measurements need also to take into account the possible variations of quality during a call. In this context, it is important to adapt duration of test calls with the possible variations of quality inside the same communication. So, it does not seen reasonable to decrease call duration below 3 minutes. The duration of test calls should allow a correct determination of voice quality stability and end to end delay stability. Therefore, 5 minutes seem to be a relevant duration for an accurate assessment. Such duration of test calls is applied only for two indicators: the listening speech quality and the end to end delay metrics (defined in clause 7 of the present document). During the call period test, these two metrics are periodically measured in order to determine stability of speech quality and variation of end to end delay at the end of the call.

10 Measurement configurations

Measurement configurations should represent the use cases of telephony services.

10.1 VoIP services

For VoIP services, there are several use cases, but the most important of them (in terms of traffic) are:

- IP to PSTN;

- IP to Mobile; and
- IP to IP.

In the context, for VoIP services, the 3 configurations IP to PSTN, IP to mobile and IP to IP have to be analyzed. It means that for every "point of measure" (define by the measurement frequency), the 3 configurations are analysed. These 3 analyses can be performed sequentially.

10.2 VoIP services in triple play context

In case of VoIP services included in triple play offers (data, VoIP, IPTV) it is necessary to assess the voice service alone and in the multiple usage condition, e.g. Voice + TV, Voice + peer-to-peer downloading, voice + TV + peer-to-peer downloading.

11 Measurement locations and their distribution

11.1 Measurement location requirements

Quality assessment of telephony service using an intrusive method (based on test calls) has the advantage to precisely analyze voice quality over defined configurations. This advantage is due to takes benefit of an end to end analysis using test signals. The whole transmission path is taken into account and the result is well correlated with user's perception.

The main drawback of this kind of analysis is to provide a limited view of the quality, because restricted to the analysed test configurations. It is known that for some kinds of transmission (e.g. packet transmission such as IP Protocol) the measurements done on one configuration are not easily applicable to the overall services.

Concerning VoIP several factors may influence telephony service quality and may introduce quality differences from a configuration to other, from a location to another one. Among these factors the enginery rules applied by the operator or service provider and the offer subscribed by the users. Other factors may be the network architecture including several equipments (home gateway, DSLAM, BAS, Call Server, Media Gateway, etc.) in the transmission path and several manufacturers for each type of equipment.

So a combination of configurations which may dramatically increase and with it may occur quality differences for the same telephony offer. It should also be indicated that the capabilities of traffic flow may change from a geographical location to another one and depending on the number of users and on network dimensions.

Only one measurement termination is not enough to obtain a macroscopic view of the service quality offered to the users. Ideally it should be necessary to deploy an analysis point per configuration but economic costs associated to probes (to buy and to install) are not realistic.

If only one analysis point and one analysis point per configuration are not satisfactory solutions for the supervision of the quality of IP telephony services, it is clear that more important the point number is better the service supervision will be. It is also important that the measurement points be located over the whole areas covered by the telephony offer but also that these measurement points be spread according to the size of "area".

In these conditions we can define 5 types of geographical areas to be covered by the analysis points:

- "Areas" with more than 1 000 000 inhabitants.
- "Areas" from 500 000 to 1 000 000 inhabitants.
- "Areas" from 250 000 to 500 000 inhabitants.
- "Areas" from 25 000 to 250 000 inhabitants.
- "Areas" with less than 25 000 inhabitants.

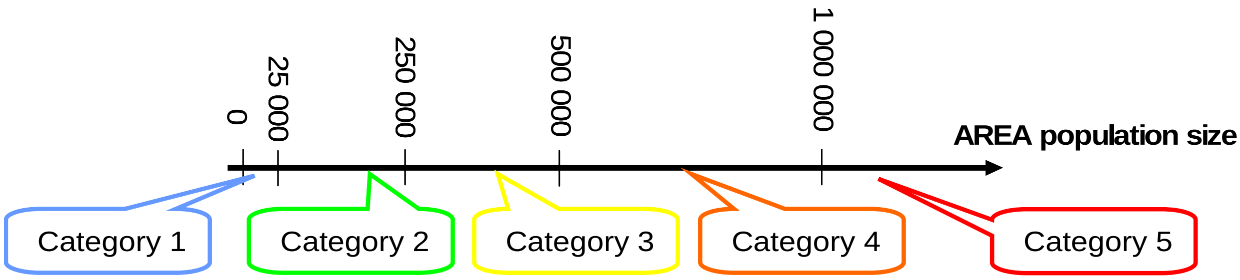


Figure 11.1: Definition of 5 "area" categories depending of population size

The term "area" can corresponds to different perimeters according to its definition. "Area" can correspond to the strict perimeter of cities or to the perimeter of a set of cities making an urban continuity. It can also correspond to a large perimeter of several cities not realizing a continuous urban pattern. The definition of "area" is specific in every country and has to take into account its characteristics (population size, urbanization mode, administration perimeters, etc.). Consequently, the definition of "area" is to be considered separately in every country.

We may also define 2 types of offer perimeters:

- Offers addressed to a potential population of users greater or equal to 10 000 000 (group 1).
- Offers addressed to a potential population of users lower than 10 000 000 (group 2).

For offers of group 1 at least 10 analysis points are to be distributed over the 5 types of geographic areas.

For offers of group 2 at least 5 analysis points are to be distributed over the 5 types of geographic areas.

11.2 Method to determine measurement locations

The following method can be applied to determine the location of the different measurement points.

For a country:

- First step consists of determining the minimum number of required of measurement point locations.
  For that purpose, it is necessary to know the potential population of users addressed by telephony offers because as described in clause 11.1, the population size determines the minimum number of analysis points required for this country. If the population of users is least than 10 millions, 5 analysis points are required. But if population of users is greater or equal to 10 millions then 10 analysis points are required.
- Second step is to define the area perimeter by taking into account the specificities of this country.
- Third step is to identify the different areas and to classify them into the 5 categories as defined above. When area classification is done, we identify the number of effective categories. According to the country size, population size, urbanization mode, and to the definition of area perimeter, it is possible that there is no area classified in category 5 or in category 1. In that case the effective number of category is lower than 5.
- Fourth step is to choose a first series of measurement points by respecting the requirement of at least one measurement location by effective category.
- Fifth step is to complete (if necessary) the location distribution with a second series of measurement points by respecting the requirement on the minimum number of measurement points (minimum number determined at the first step). In step 4 and 5, it is recommended to choose the location of analysis points by respecting a rather homogeneous geographical distribution.

To illustrate this method of choice and allocation of measurement points, two examples are presented in annex D.

12 Results presentation

For reporting, the presentation of each metric will be made separately. Each metric will be presented on its own scale.

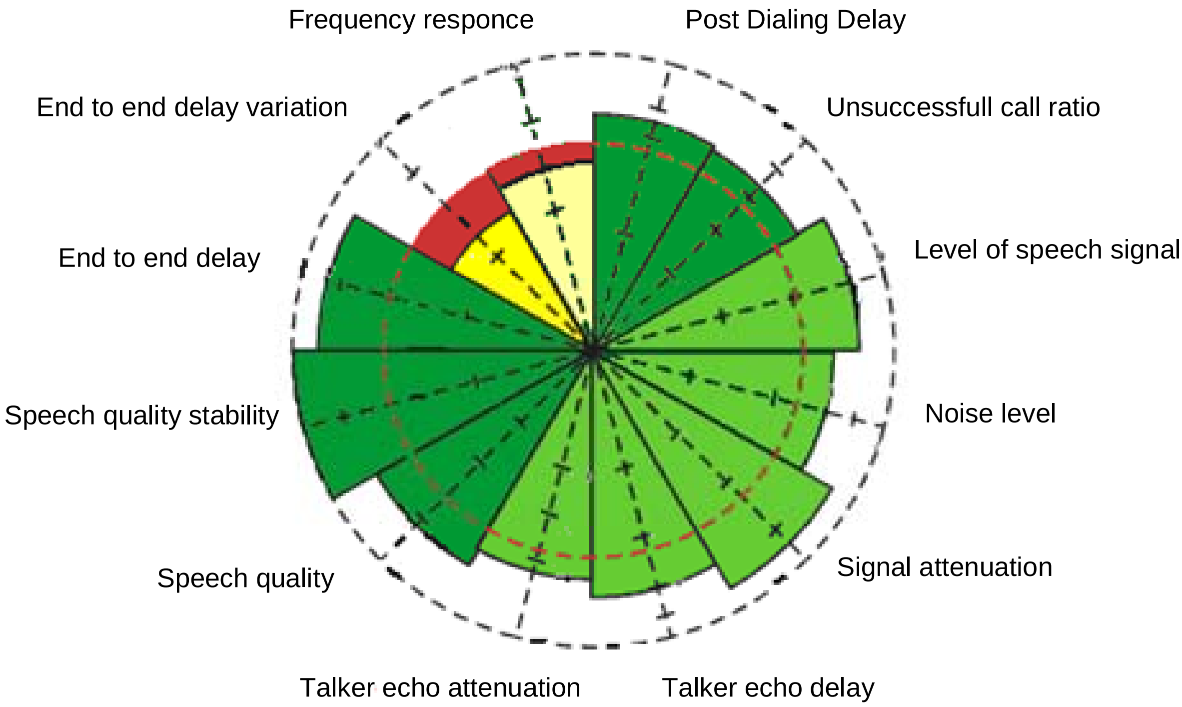
For each indicator, presentation results consist at least of presenting metric value, number of measurements to determine the metric value and measurement standard deviation.

But in order to give a quick overview of all quality parameters, a specific representation (overview visualization defined in ITU-T Recommendation P.505 [i.15]) of the metric value will be used. This representation reveals at one glance the strengths and weaknesses of telephony services by reference to non-compliant limits.

12.1 One-view visualization of performances

The one-view visualization is based on a circular presentation of indicators ("pie diagram") where each metric value represents a circle segment of the diagram.

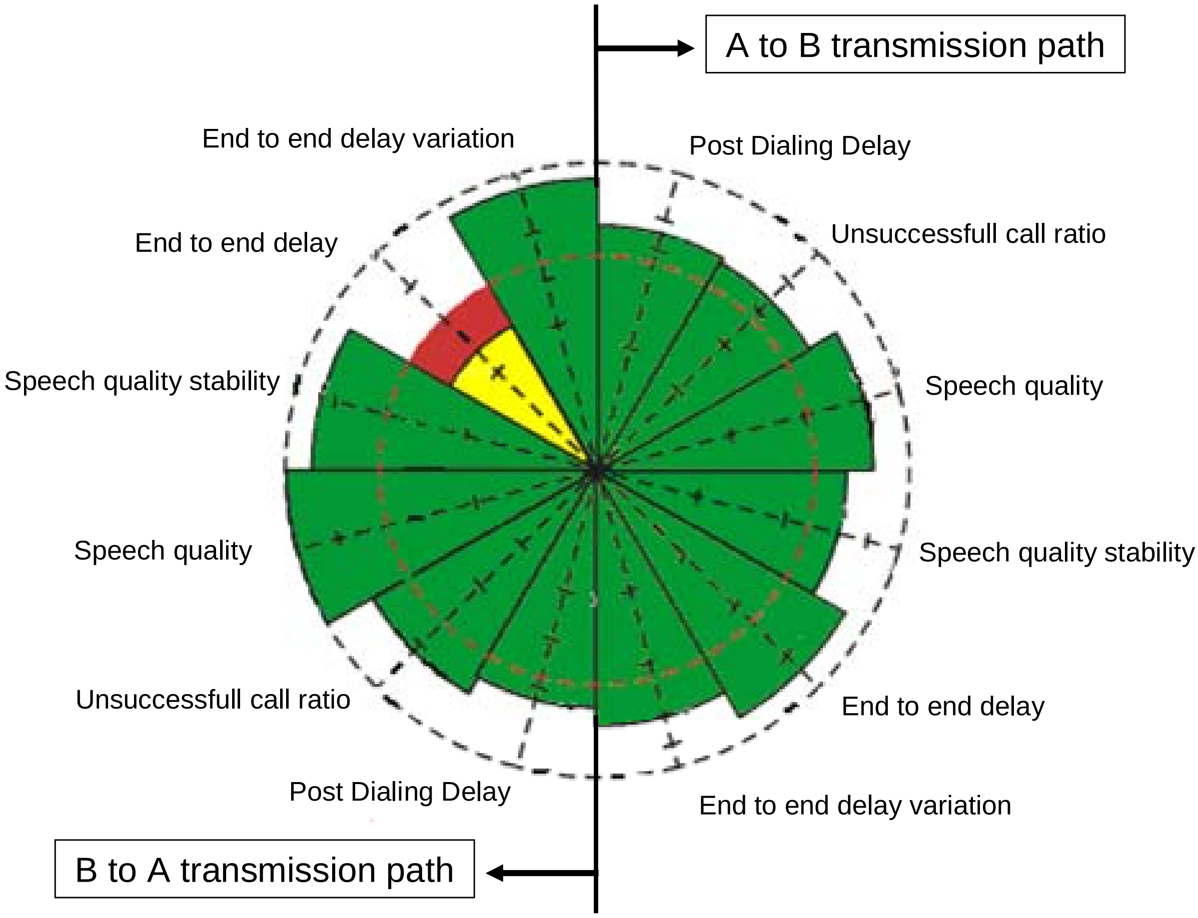
12.1.1 Pie diagram with all indicators



NOTE: This representation is associated to one of the transmission paths.

Figure 12.1: Example of "pie diagram" representation performed with all the indicators (mandatory and optional)

12.1.2 Pie diagram with mandatory indicators



NOTE: This representation presents performances in both transmission paths.

Figure 12.2: Example of "pie diagram" representation performed with only mandatory the indicators

12.2 Non-compliant limits for result visualization

Table 12.1

Indicator name	Value of non-compliant limit	Unit
Post Dialling Delay	6 000	Millisecond
Media establishment delay	1 000	Millisecond
Unsuccessful call ratio	2	%
Premature release probability		%
Level of active speech signal		dBm
Noise level		dBm0p
Noise to signal ratio		dB
Speech signal attenuation		dB
Echo-Annoyance factor		
Listening speech quality		
Listening speech quality stability		
End to end delay	200	Millisecond
End to end delay variation		
Frequency response		

NOTE: When we will have elements allowing to fix new compliant limit, table 12.1 will be amended.

13 Publication of the results

For the supervision of a telephony offer, a publication of the characterization results of this offer can be published in a monthly, weekly even daily way.

For the supervision and the comparison of several telephony offers, a publication of the characterization results of these offers can be published in a monthly and/or weekly way. A daily publication is not adapted to these particular conditions because it corresponds to an analysis window too short giving no real sense to the comparisons.

Annex A (normative): Indicator stability formulation

A.1 Presentation

The purpose is to assess stability indicator for different metrics like listen speech quality and end to end delay. These stability indicators should characterize metric stability during the call.

For speech quality, the stability indicator should be independent of the intrinsic quality of the codec because stability indicator completes speech quality indicator. So, stability indicator is base on derived formulation. It is "simple" formulation, easy to be implemented.

It is also a dynamic indicator, i.e. it can be determined during the call and does not need that communication is ended to begin the determination.

A.2 Formulation

The assessment of METRIC Stability (ST-METRIC) may be defined in 5 steps:

1) To measure the METRIC periodically over the duration of one communication. N measurements provide N METRIC values.

2) For each METRIC (i) value i from 2 to N, METRIC_GAP(i) is calculated as the absolute difference with the previous value METRIC (i-1):

   METRIC_GAP (i) =|METRIC (i)- METRIC (i-1)|

3) In order to take into account the subjective perception and measurement accuracy, METRIC_GAP (i) is set to 0 when the difference is equal or lower to THRESHOLD1:

   if METRIC_GAP (i) > (2 x THRESHOLD1), then METRIC_GAP (i) = METRIC_GAP (i);

   if THRESHOLD1 < METRIC_GAP (i) ≦ 2 x THRESHOLD1, then METRIC_GAP (i) = [METRIC_GAP (i) x 2] - (2 x THRESHOLD1);

   if METRIC_GAP (i) ≦ THRESHOLD1, then METRIC_GAP (i) = 0.

4) The instability (INS_METRIC) associated to the METRIC over the whole N measurements is defined by mean value of METRIC_GAP (i).

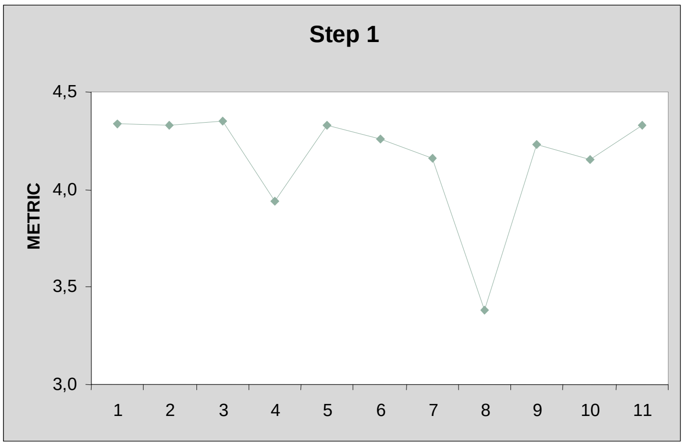
   INS_METRIC = 1/(N-1) Σ METRIC_GAP (i) with i = [2:N].

5) A linear weighting function is applied in order to express Stability ST-METRIC on a 0 to 100 scale.

This formulation can be used to determine Listen Speech Quality Stability (ST-MOS) and end to end delay Stability (ST-Delay).

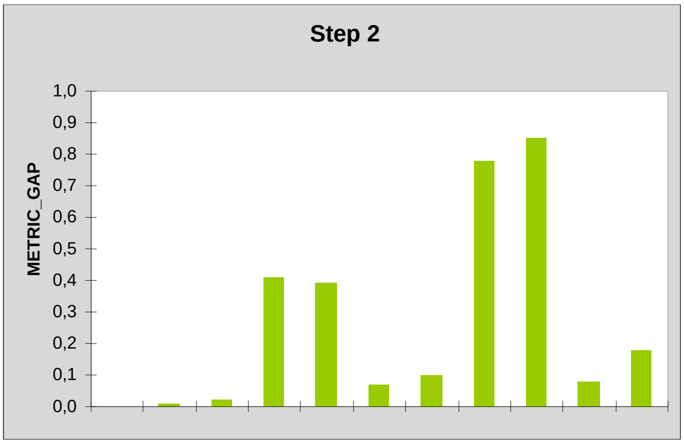
For these two metrics, only THRESHOLD1 value and linear weighting function are specifics.

A.3 Graphic illustration of the formulation



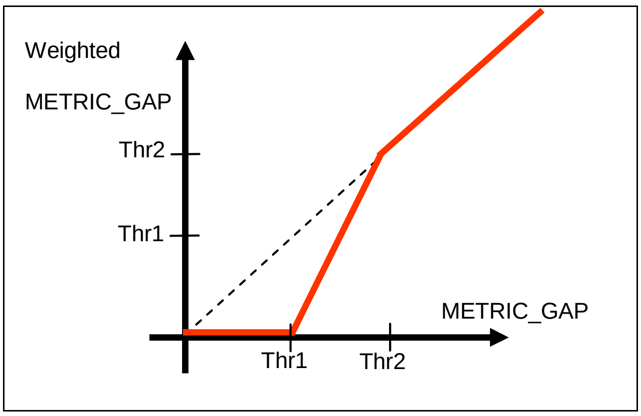
NOTE: N METRIC value is determined during a call.

Figure A.1



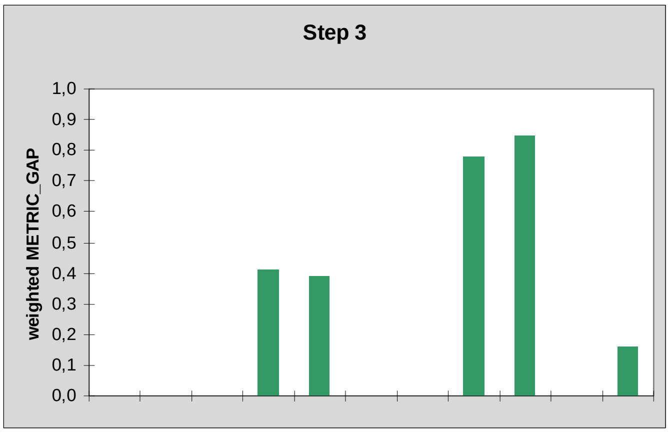
NOTE: For each METRIC value the difference with the previous value is calculated.

Figure A.2



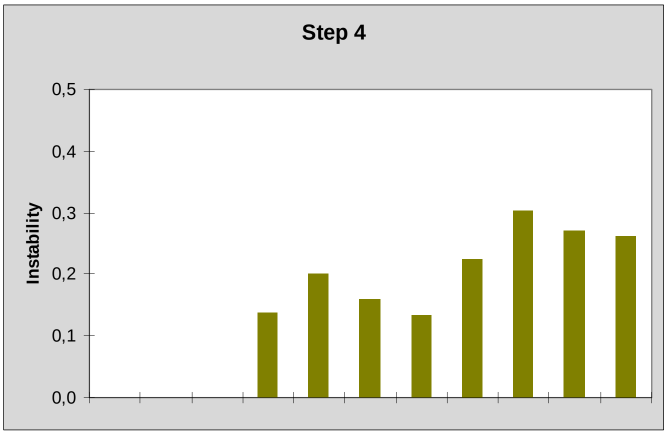
NOTE: The weighted function applies to METRIC_GAP values:
Thr1 = THRESHOLD1 and Thr2 = 2 x THRESHOLD1.

Figure A.3



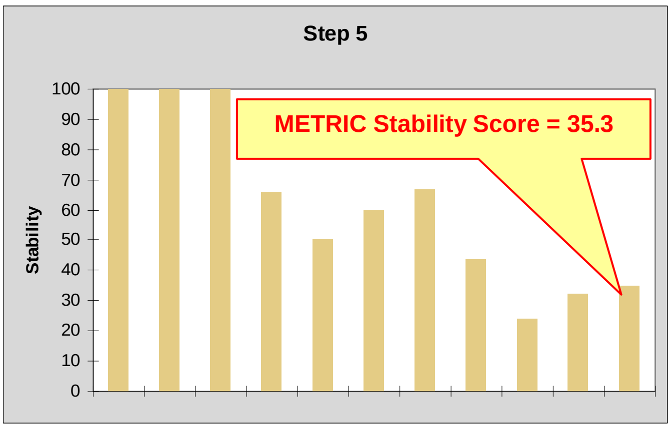
NOTE: Weighted METRIC_GAP calculated with the weighted function.

Figure A.4



NOTE: The METRIC instability is calculated for each METRIC_GAP value.

Figure A.5



NOTE: Stability score is calculated for each instability value and the last value represents the METRIC stability score for the call.

Figure A.6

A.4 Some examples of stability indicator calculated on Listening Speech Quality

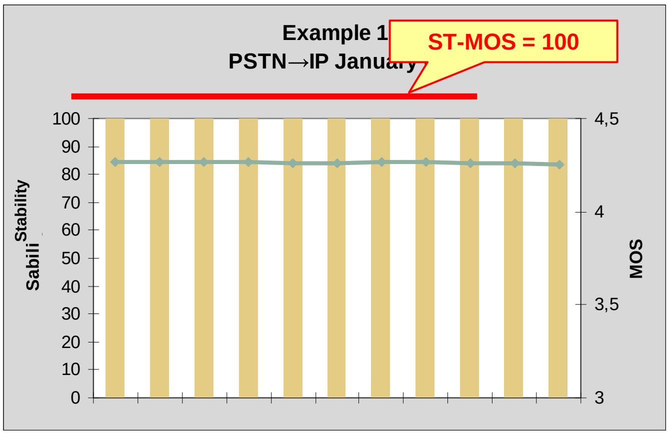


Figure A.7

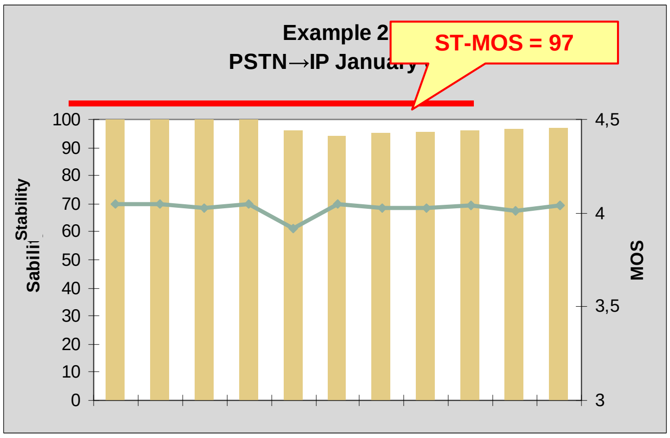


Figure A.8

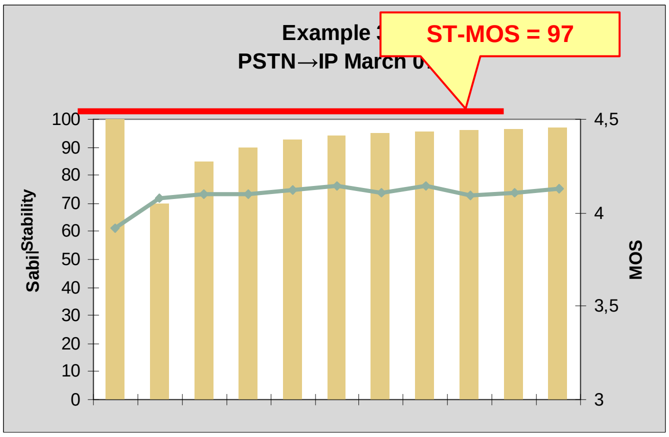


Figure A.9

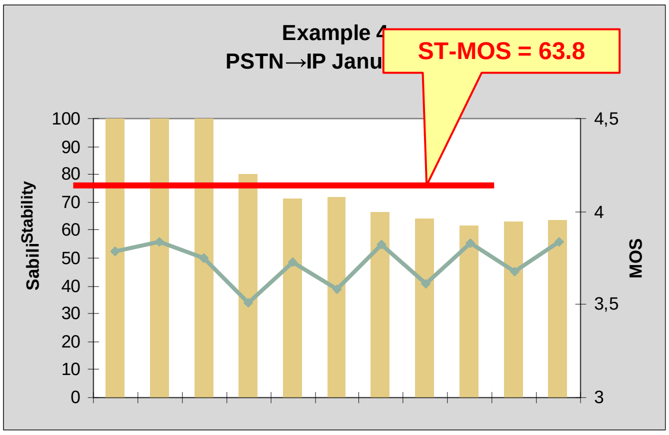


Figure A.10

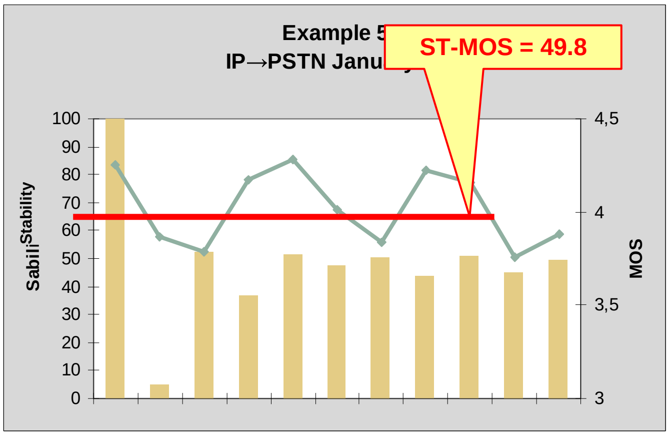


Figure A.11

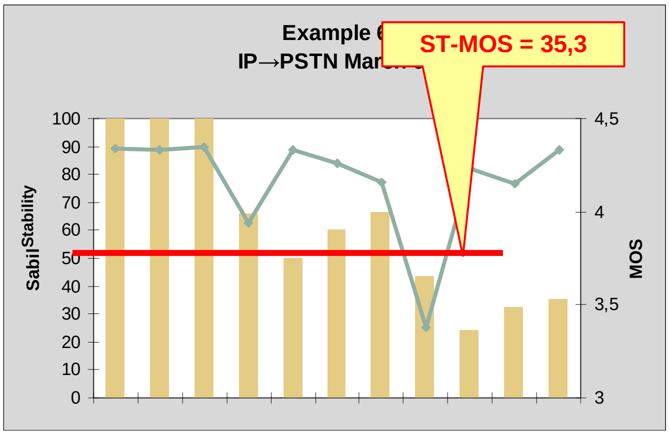


Figure A.12

Annex B (normative): Calibration to take into account the frequency response of transducers

Sending direction

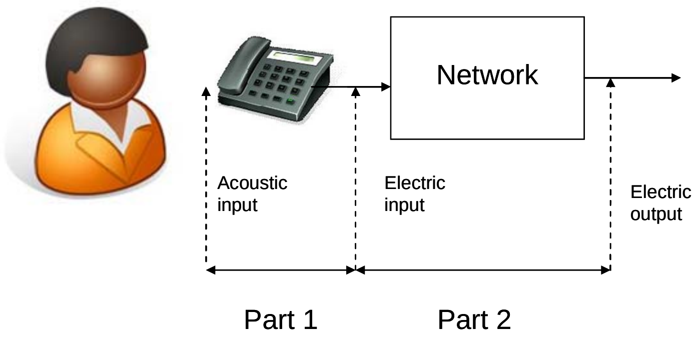


Figure B.1

In most cases quality evaluation is performed for part 2 (electric to electric), but it is necessary to take into account the influence of part 1 (acoustic to electric).

Measurements taking into account acoustic are more complex (necessity of quiet environment, defined stimuli, etc.)

Receiving direction

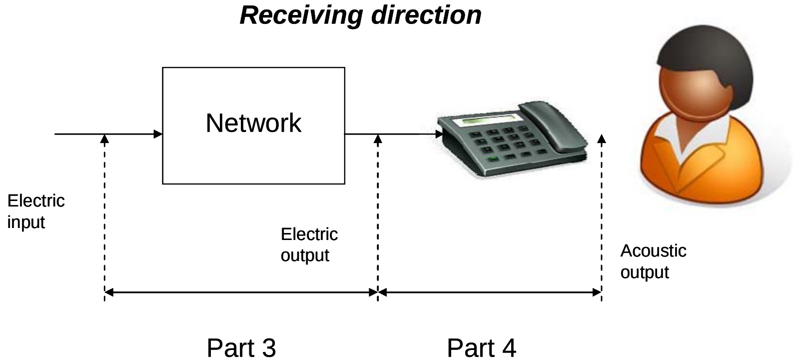


Figure B.2

For receiving, the problem is the same, electric to acoustic (part 4) has to be taken into account, with the same difficulties as for sending.

B.1 Method presentation

B.1.1 Sending

Make a transfer function for part 1. Artificial voice (ITU-Recommendation P.50 [i.19] or ITU-Recommendation P.501 [i.20] signal) can be used or any desired signal with HATS as artificial mouth.

For measurement of part 2 two options:

- Resulting signal can be used for analyzing network.
- Transfer function obtained can be inserted in measurement (transfer function used here is electric to electric taking into account HATS).

B.1.2 Sending

Make a transfer function for part 4. Artificial voice (ITU-Recommendation P.50 [i.19] or ITU-Recommendation P.501 [i.20] signal) can be used or any desired signal (with HATS as measuring system).

The obtained transfer function is applied to the result of measurement of part 3.

B.1.3 Global communication

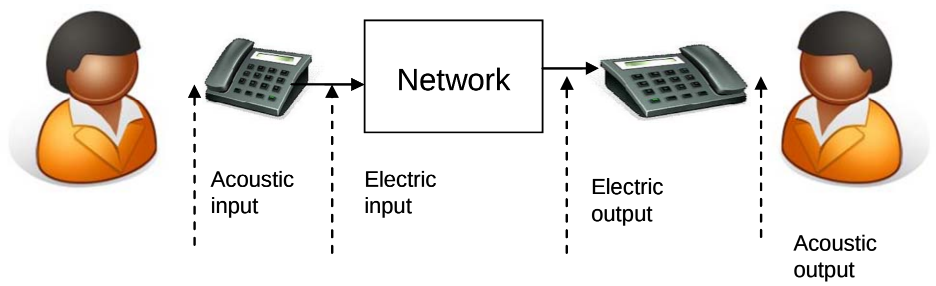


Figure B.3

First transfer function (sending part of terminal) will be applied to signal at input of network and second transfer function (receiving part of terminal) will be applied to resulting signal.

B.1.4 Applications

Network with defined terminals:

- Transfer functions are measured at the beginning for the terminal and all following measurements can be made without taking into account the acoustic part.

Network using unknown terminals:

- A first characterization can be made with transfer function corresponding to a standard (average) terminal.

Annex C (informative): Echo presentation

C.1 Talker echo

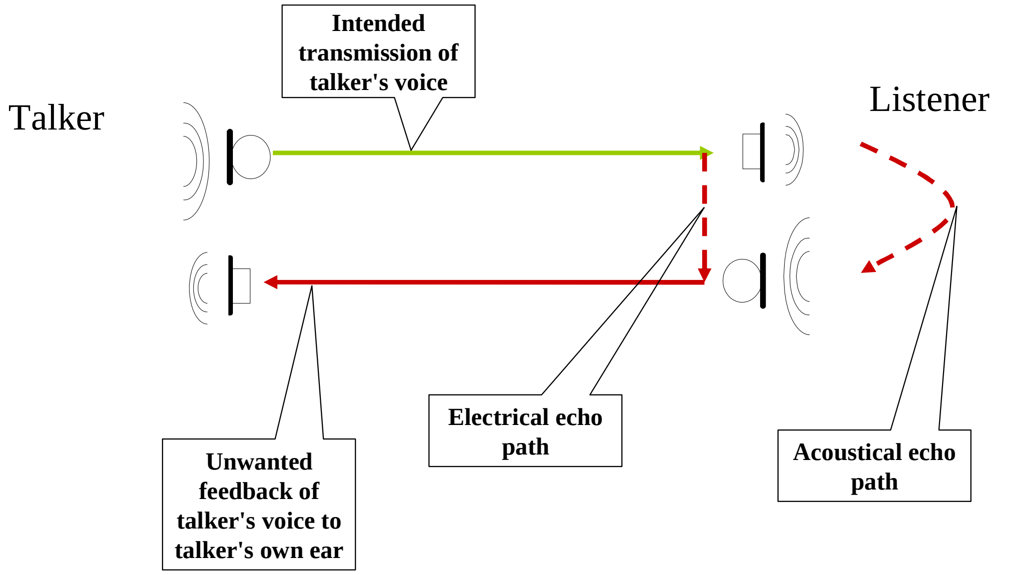


Figure C.1

C.2 Listener echo

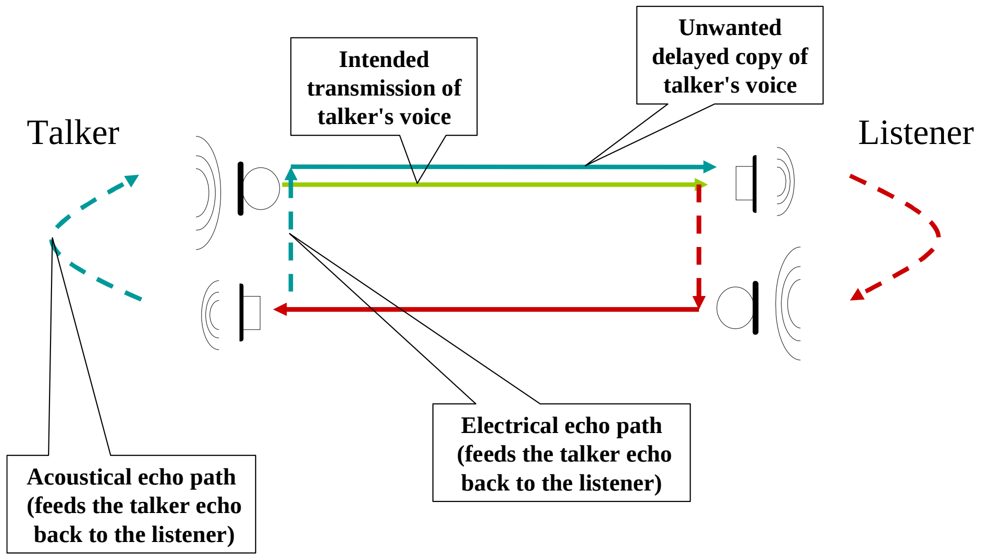


Figure C.2

Annex D (informative): Examples of measurement point distribution

In this annex, two hypothetical examples are presented to illustrate the method of choice and allocation of measurement points described in clause 11.2.

D.1 Example of France

Step 1: Determination of the minimum number required of measurement point locations

Population size in France = 63 800 000. In this context, supervision of telephony offers needs at least 10 measurement point locations.

Step 2: Definition of area perimeter

Considering the topology of the urbanization in France, it makes sense to define the term area as perimeter of a set of cities making an urban continuity.

Step 3: Identification of the different areas and classification into the 5 area categories defined in clause 11.1

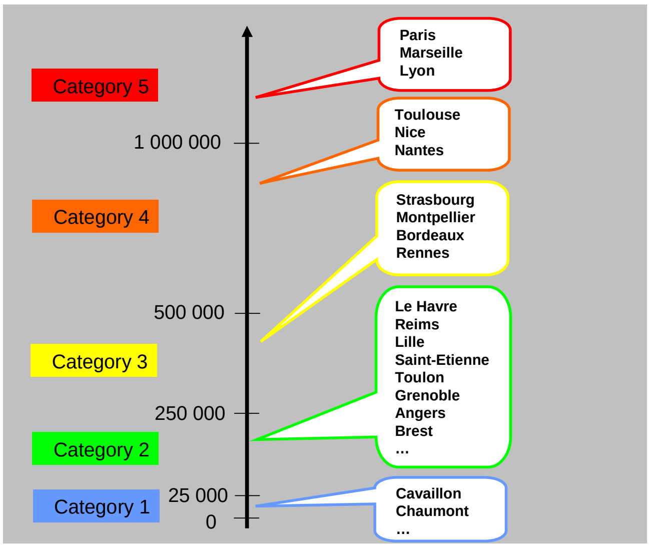


Figure D.1

We notice for France 5 effective area categories

Step 4: Choice of a first series of measurement points

Respecting the requirement of at least one measurement location by effective category, the first series of measurement points can be this one:

- Category 5: Paris.
- Category 4: Toulouse.
- Category 3: Bordeaux.
- Category 2: Dijon.
- Category 1: Lannion.

This first identification is realized by distributing geographically the measurement points.

Step 5: Complete the location distribution

5 others measurement points are defined to respect the requirement of at least 10 measurement point locations is needed for supervision of telephony services in France. The second series of measurement points can be this one:

- Category 5: Marseille, Lyon.
- Category 4: Nantes.
- Category 3: Strasbourg.
- Category 2: Lille.

This second identification is realized by distributing geographically the measurement points and also by taking into account that the French population lives mainly in city (75 %). In these conditions, two areas in category 5 are preferred to another measurement point in category 1.

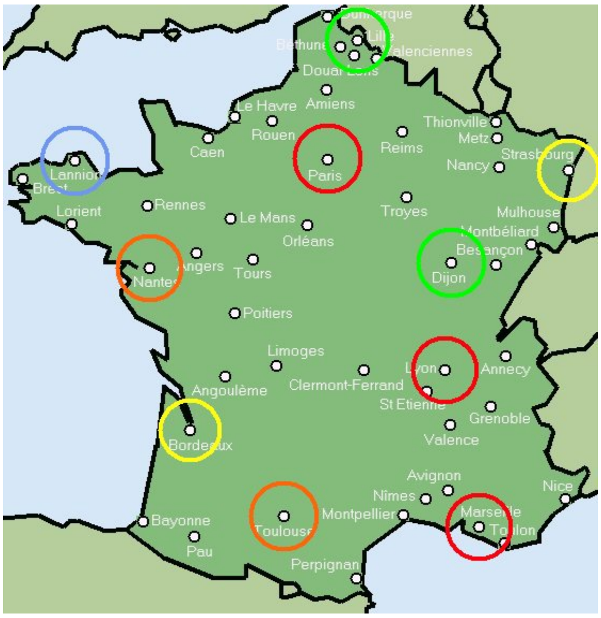


Figure D.2

D.2 Example of Switzerland

Step 1: Determination of the minimum number required of measurement point locations

Population size in Switzerland = 7 590 000. In this context, supervision of telephony offers needs at least 5 measurement point locations.

Step 2: Definition of area perimeter

It makes sense to define the term area as perimeter of cities.

Step 3: Identification of the different areas and classification into the 5 area categories defined in clause 11.1

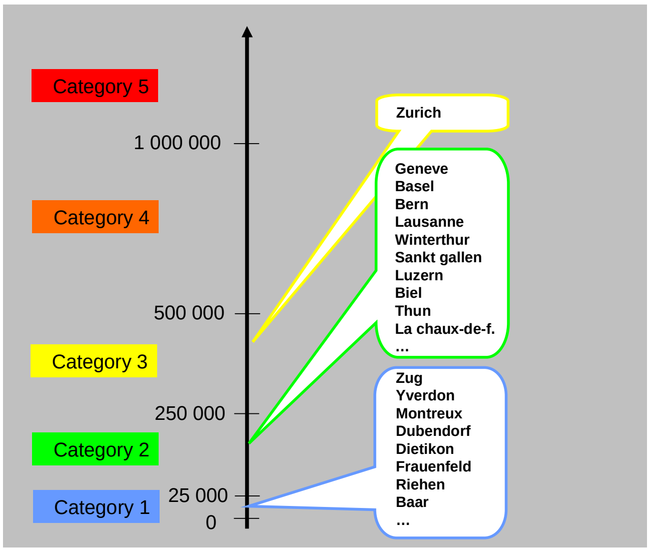


Figure D.3

We notice for Switzerland 3 effective area categories.

Step 4: Choice of a first series of measurement points

Respecting the requirement of at least one measurement location by effective category, the first series of measurement points can be this one:

- Category 3: Zurich.
- Category 2: Geneva.
- Category 1: Locarno.

This first identification is realized by distributing geographically the measurement points.

Step 5: Complete the location distribution

Two others measurement points are defined to respect the requirement of at least 5 measurement point locations is needed for supervision of telephony services in Switzerland The second series of measurement points can be this one:

- Category 2: Bern.
- Category 1: Altdorf.

This second identification is realized by distributing geographically the measurement points and also by taking into account the Swiss area size (lower than 250 000 thus classified in categories 1 and 2). The choice of the measurement points (realized in steps 4 and 5) also takes into account a specificity of Switzerland: 3 regions associated with different languages (French, German and Italian).

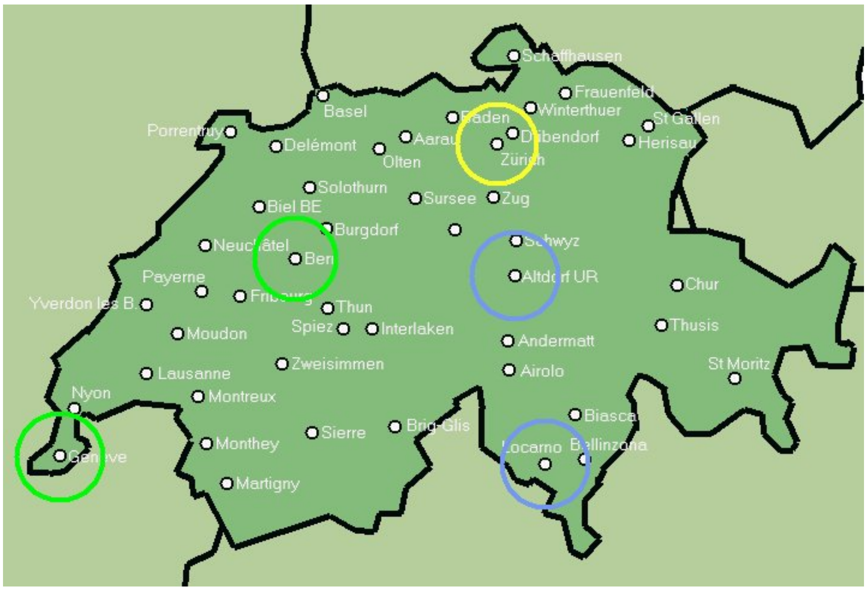

Basel
Porrentruy
Delémont
Aarau
Olten
Winterthuer
Frauenfeld
St Gallen
Herisau
Solothurn
Sursee
Zug
Biel BE
Burgdorf
Bern
Thun
Spiez
Interlaken
Payerne
Yverdon les B.
Moudon
Zweisimmen
Lausanne
Nyon
Montreux
Monthey
Sierre
Martigny
Andermatt
Airolo
Chur
Thusis
St Moritz
Biasca
Altdorf UR

Figure D.4

History

Document history		
V1.1.1	February 2009	Publication as EG 202 765-2
V1.1.3	March 2010	Membership Approval Procedure MV 20100507: 2010-03-08 to 2010-05-07