

International Telecommunication Union

ITU-T

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

P.806

(02/2014)

SERIES P: TERMINALS AND SUBJECTIVE AND OBJECTIVE ASSESSMENT METHODS

Methods for objective and subjective assessment of
speech quality

A subjective quality test methodology using multiple rating scales

Recommendation ITU-T P.806

ITU-T P.806 (02/2014)

ITU-T

ITU-T P-SERIES RECOMMENDATIONS

TERMINALS AND SUBJECTIVE AND OBJECTIVE ASSESSMENT METHODS

Vocabulary and effects of transmission parameters on customer opinion of transmission quality	Series	P.10
Voice terminal characteristics	Series	P.30
		P.300
Reference systems	Series	P.40
Objective measuring apparatus	Series	P.50
		P.500
Objective electro-acoustical measurements	Series	P.60
Measurements related to speech loudness	Series	P.70
Methods for objective and subjective assessment of speech quality	Series	P.80
		P.800
Audiovisual quality in multimedia services	Series	P.900
Transmission performance and QoS aspects of IP end-points	Series	P.1000
Communications involving vehicles	Series	P.1100
Models and tools for quality assessment of streamed media	Series	P.1200
Telemeeting assessment	Series	P.1300
Statistical analysis, evaluation and reporting guidelines of quality measurements	Series	P.1400
Methods for objective and subjective assessment of quality of services other than voice services	Series	P.1500

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

Recommendation ITU-T P.806

A subjective quality test methodology using multiple rating scales

Summary

Recommendation ITU-T P.806 describes a methodology for evaluating the subjective quality of speech samples using multiple rating scales. In addition to scores for overall quality and loudness, the methodology yields scores for six perceptual quality (PQ) attributes of the speech sample. Each of these PQ scores is based on ratings of the amount or degree of degradation present in the sample for an attribute that underlies listener's judgment of speech quality. Four of the PQ scores represent degradation associated with the speech signal and two of the PQ scores represent degradation associated with the background noise. The methodology is designed to be used with naive subjects and yields scores for overall quality and loudness plus scores for the six PQ attributes of the speech sample. These PQ scores can be used to provide diagnostic information on the underlying causes of speech quality degradation.

This Recommendation includes an electronic attachment containing audio samples for the conditions described in Appendix I.

History

Edition	Recommendation	Approval	Study Group	Unique ID*
1.0	ITU-T P.806	2014-02-13	12	11.1002/1000/12125

Keywords

Diagnostic evaluation of speech quality, multi-dimensional quality assessment, speech quality evaluation, subjective testing.

* 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

In this Recommendation, the expression "Administration" is used for conciseness to indicate both a telecommunication administration and a recognized operating agency.

Compliance with this Recommendation is voluntary. However, the Recommendation may contain certain mandatory provisions (to ensure, e.g., interoperability or applicability) and compliance with the Recommendation is achieved when all of these mandatory provisions are met. The words "shall" or some other obligatory language such as "must" and the negative equivalents are used to express requirements. The use of such words does not suggest that compliance with the Recommendation is required of any party.

INTELLECTUAL PROPERTY RIGHTS

ITU draws attention to the possibility that the practice or implementation of this Recommendation 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 Recommendation development process.

As of the date of approval of this Recommendation, ITU had not received notice of intellectual property, protected by patents, which may be required to implement this Recommendation. 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/>.

© ITU 2014

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

Table of Contents

	Page
1 Scope.....	1
2 References.....	1
3 Definitions	1
3.1 Terms defined elsewhere	1
3.2 Terms defined in this Recommendation.....	1
4 Abbreviations and acronyms	2
5 Conventions	2
6 Methods and procedures	2
6.1 User interface.....	4
6.2 Listener instructions and training	5
6.3 Designing an ITU-T P.806 test.....	5
6.4 Organization of an ITU-T P.806 test session	6
7 ITU-T P.806 test results.....	7
Appendix I – ITU-T P.806 test instructions in English	10
Appendix II – ITU-T P.806 test instructions in French	13
Bibliography.....	16

Electronic attachment containing audio samples for the conditions described in Appendix I.

Introduction

In most standard ITU-T subjective test methodologies for evaluating speech quality, subjects are passive participants in the exercise. Typically, subjects listen to the test sample and provide a judgement of the overall quality of recorded passages of speech materials. These listening quality test methodologies involve a single rating scale and the quality estimate is an average of the ratings for multiple subjects where each subject typically rates samples from multiple talkers. [ITU-T P.800] describes a number of such methodologies including the absolute category rating (ACR) method, which produces the mean opinion score (MOS).

More than three decades ago, the diagnostic acceptability measure (DAM) [b-Voiers] was developed to evaluate the underlying causes of degradation in speech quality. The DAM used 21 rating scales, nine associated with degradation in the speech signal alone, eight associated with degradation in the background noise alone, and four associated with overall quality. The MOS rating scale was one of those four. The DAM required expert subjects who were screened, trained and calibrated to provide reliable and consistent responses on the large number of rating scales utilized by the DAM. While the DAM enjoyed considerable success for evaluating speech quality in government and industry in the United States, it was a proprietary method and not suited for routine testing with naive subjects.

[ITU-T P.835] was developed to evaluate speech quality under conditions of noise suppression. In the [ITU-T P.835] method, naive subjects evaluate each sample in two dimensions using rating scales to estimate the amount of distortion in the speech signal and the degree of intrusiveness of the background noise, before making their rating of overall quality (OVRL). The process of evaluating the sample separately on speech distortion and background intrusiveness, conditions the subjects to integrate the effects of both sources of degradation in making their ratings of overall quality. Routine use of the [ITU-T P.835] test methodology has shown that naive subjects can effectively and reliably use multiple rating scales to evaluate the quality of speech in background noise. The ITU-T P.806 test methodology extends the multiple rating scale approach of [ITU-T P.835] to the more general case of speech in most types of degradation.

Recommendation ITU-T P.806

A subjective quality test methodology using multiple rating scales

1 Scope

In this Recommendation¹, a subjective test methodology has been developed as a general subjective test for evaluating speech quality in "listening only" test scenarios. This Recommendation is appropriate for use in a wide variety of bandwidths, including clean and error channel conditions, and clean and noisy background conditions. However, as opposed to most "listening only" tests (for example, [ITU-T P.800] test variants and [ITU-T P.835]), such different categories of test conditions can be mixed within the same ITU-T P.806 test and still provide meaningful results. Results derived from ITU-T P.806 testing are less susceptible to the effects of test context (i.e., the overall composition of the conditions within a test) than any other Recommendation for "listening only" subjective speech quality testing.

2 References

The following ITU-T Recommendations and other references contain provisions which, through reference in this text, constitute provisions of this Recommendation. At the time of publication, the editions indicated were valid. All Recommendations and other references are subject to revision; users of this Recommendation are therefore encouraged to investigate the possibility of applying the most recent edition of the Recommendations and other references listed below. A list of the currently valid ITU-T Recommendations is regularly published. The reference to a document within this Recommendation does not give it, as a stand-alone document, the status of a Recommendation.

- [ITU-T P.800] Recommendation ITU-T P.800 (1996), *Methods for subjective determination of transmission quality*.
- [ITU-T P.835] Recommendation ITU-T P.835 (2003), *Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm*.

3 Definitions

3.1 Terms defined elsewhere

None.

3.2 Terms defined in this Recommendation

This Recommendation defines the following terms:

- 3.2.1 background-level (B-LVL):** Degradation due to the level of the background noise, background that may be described as hissing, rushing, roaring.
- 3.2.2 background-variability (B-VAR):** Degradation due to the variability or non-stationarity of the background noise, background that may be described as bubbling, intermittent, variable.
- 3.2.3 LOUD:** Overall loudness of the combination of speech signal and background noise.
- 3.2.4 OVRL:** Overall quality of the combination of speech signal and background noise.

¹ This Recommendation includes an electronic attachment containing audio samples for the conditions described in Appendix I.

3.2.5 signal-fluttering (S-FLT): Slow-varying degradation in the speech signal, speech that may be described as fluttering, babbling, discontinuous.

3.2.6 signal-high-frequency coloration (S-HFC): Degradation in the lower end of the spectrum of the speech signal, speech that may be described as small, distant, thin.

3.2.7 signal-low-frequency coloration (S-LFC): Degradation in the higher end of the spectrum of the speech signal, speech that may be described as dull, muffled, smothered.

3.2.8 signal-rough (S-RUF): Fast-varying degradation in the speech signal, speech that may be described as rough, raspy, harsh.

4 Abbreviations and acronyms

This Recommendation uses the following abbreviations and acronyms:

ACR	Absolute Category Rating
DAM	Diagnostic Acceptability Measure
MOS	Mean Opinion Score
PQ	Perceptual Quality

5 Conventions

"ITU-T P.806" is used as the descriptive name for the subjective test described in this Recommendation.

6 Methods and procedures

Methods for the subjective evaluation of speech quality in transmission systems and equipment have typically involved the use of a single rating scale, specifically one of the rating scales described in [ITU-T P.800]. The most widely used scale is the mean opinion score (MOS) described in [ITU-T P.800] for use in the absolute category rating (ACR) methodology. In a typical ACR test, subjects listen to a two-sentence, 8-s sample of speech and then rate the sample using the five-category MOS rating scale. The categories are labelled 'Bad', 'Poor', 'Fair', 'Good', and 'Excellent', which correspond to ratings 1, 2, 3, 4 and 5, respectively. The ACR method and the MOS scale are designed to be used with naive subjects as defined in [ITU-T P.800]. For each test condition, the MOS is computed as the average of the individual ratings over talkers and over subjects. The resulting value of MOS is a summary estimate of speech quality resulting from the combined judgments of multiple subjects each of whom may be judging the degradation on different aspects of the speech signal itself, the background, or a combination of the two. Each subject may describe that degradation in vastly different terms, but there are a limited number of independent factors that underlie such subjective judgements of speech quality. Identification of the underlying factors is usually accomplished through multivariate statistical analysis. The nature and number of these underlying factors will vary depending on the specific statistical technique used, as well as the size and variety of the data from which they are derived. For this Recommendation, a database of subjective ratings on more than 20 rating scales for more than 1'000 test conditions has been analysed using principal components analysis [b-AH-11-028], [b-C228], [b-Sen]. The results from this analysis have led to the identification of six perceptual quality (PQ) rating scales that underlie subjects' judgments of speech quality. For each of these PQs, a set of common descriptive terms has been compiled. Subjects can use these terms to indicate how much of a given PQ is present in a speech sample. Four of the PQs refer to degradation in the quality of the speech signal and two of the PQs refer to degradation in the quality of the background. Table 6-1 shows a list of the six PQs along with the set of descriptive terms for each PQ.